A COMPARATIVE STUDY OF ROBUSTNESS OF ZIPF'S LAW ACROSS LITERATURES

Thesis submitted to the

BUNDELKHAND UNIVERSITY, JHANSI
For the Award of the Degree of
DOCTOR OF PHILOSOPHY
In
LIBRARY AND INFORMATION SCIENCE

By
(Ms.) Monika Jauhari

Under the supervision of :

Guide :
Prof. J.N. Gautam
Dean, Faculty of Arts
School of Studies in Library &
Information Science,
Jiwaji University, Gwalior

Co-Guide :
Prof. M.T.M. Khan
Dean, Faculty of Arts and Head
Department of Library &
Information Science,
Bundelkhand University, Jhansi

Department of Library and Information Science
BUNDELKHAND UNIVERSITY, JHANSI (U.P.)
2007

Contents

CERTIFICATE

This is to certify that the work embodied in the thesis entitled, "A comparative study of robustness of Zipf's Law across literatures" is submitted by Monika Jauhari for the award of degree of Doctor of Philosophy in Library & Information Science. It is a record of bonafide research work carried out by her under our supervision and guidance. This work has not been submitted elsewhere for a degree or a diploma in any form.

It is further certified that she has worked with us for the period required for the Ph.D. degree and has followed, as far as we know, relevant directions laid down in the ordinance for Ph.D. degree of the Bundelkhand University.

GUIDE

Prof. J.N.Gautam,
Dean, Faculty of Arts,
School of Studies in Library &
Information Science,
Jiwaji University, Gwalior

CO-GUIDE

Prof. M.T.M. Khan,
Dean Faculty of Arts and
Head, Dept. of Library &
Information Science,
Bundelkhand University, Jhansi

Dated: 10/8/07

DECLARATION

I hereby declare that the thesis entitled "A comparative study of robustness of Zipf's Law across literatures" is my own work conducted under the joint supervision of Prof. J. N. Gautam (School of Studies in Library & Information Science, Jiwaji University, Gwalior) and Prof. M. T. M. Khan (Head, Dept. of Library & Information Science, Bundelkhand University, Jhansi), as was approved by the research degree committee. I have put in the required period of attendance with the supervisor at the centre.

I further declare that to the best of my knowledge, the thesis does not contain any part of any work, which has been submitted for the award of any degree in the university or any other university.

Monika Jauhari

Monika Jauhari
Student

Dated: 10·08·07

Preface

In the field of Linguistics, Bibliometrics & Informetrics, Zipf's law commends a high influence. There are numerous recent applications of Zipf's law in Linguistics, Internet research, Geography, Medicine and Economics. Many people call Zipf's law as "one of the most puzzling phenomena in bibliometrics". Zipf's Law approximates the relationship between rank and frequency of any text. It describes the fact that when words are ranked on frequency, from most to least frequent, plotting rank against frequency yields a hyperbolic curve.

Zipf attributed this law as a consequence of "Principle of Least Effort". Principle of Least Effort is relevant even today. If one views Zipf's law in terms of communication costs- one infers that communication costs increase as the number of words and their length grows. Thus, Zipf's law is applicable in understanding human language. There have been many applications of the law in natural languages, like Hindi, English, Urdu, Irish, Latin, Vietnamese, Chinese, Russian, Voyanich manuscript and random texts etc.

Zipf's goal was to put language study on a par with exact sciences, by use of "statistical techniques". Zipf attempted to prove that the key to the explanation of all synchronic and diachronic language-phenomena has been found in a statistically estimated tendency to maintain equilibrium between size and frequency. It is said that Zipf's law is perhaps the best-

known model of word probabilities. Researchers argued that Zipf's law is a reflection of a specific property of the organization of human memory, which usually operates with more frequent language units in all cases of the spontaneous use of speech.

The present work postulates the hypotheses that the principle of least effort is a universal phenomenon; all writers would follow an economy in the use of words irrespective of the language concerned and the rank-frequency distribution of words would be similar in all languages and aims to attain the objective of finding the interrelationships between the rank and the frequency of a word in selected literatures; test the applicability of Zipf's law in diverse literatures and compare this applicability.

The thesis is divided in the following chapters: -

Chapter 1 introduces Zipf's law by defining and illustrating Mathematical Foundations of Zipf's Law. It highlights the major thoughts on the Zipf's Law held by various researchers such as Hill, Price, Mandelbrot, Herdan and Haitun. It also convoluted many more approaches on Zipf's Law.

Chapter 2 does a review of literature on the applications of Zipf's Law. It illustrates the application of Zipf's law in city populations, growth pattern of production companies, features of the Internet, finance and business, firm sizes, ecological systems, genomic data, earthquakes and clinical diagnosis etc. It also describes how apart from aforesaid languages, Zipf's

law in applicable in technical subjects, random texts, Monkey-type texts and legal texts also.

Chapter 3 discusses the research methodology involved in this work by defining the objectives, hypothesis and the data used. It filters out the reason for choosing appropriate software and describes the ones that are used in this work. It also discusses the ranking methods involved in ranking the word frequencies. A note on nonlinear regression illustrates the models involved. It defines the model families and illustrates their types. A note on Project Gutenberg e-texts and the IIT Kanpur's e-text is also embedded.

Chapter 4 presents the analysis of first sets of documents viz. Computer Science Literature, English Literature, German Literature, Zipf's Law for English-German Business Dictionary, Hindi Literature, Library Science Literature, Urdu Literature and Sanskrit Literature. It also relates Zipf's law and Flesch Readability Index. An effort is made in this chapter to highlight principle of least effort.

Chapter 5 discusses issues pertaining to interrelationships between the rank & frequency; issues related to robustness of Zipf's Law and issues related to inter-literature comparison of the applicability of Zipf's law. Chapter 6 presents summary, conclusions and suggestions for future research.

ACKNOWLEDGEMENTS

"Somehow I can't believe that there are many heights that can't be scaled by a man who knows the secret of making dreams come true. This special secret can be summarized in four C's. They are Curiosity, Confidence, Courage and Constancy and the greatest of these is CONFIDENCE".

First of all, I want to honor the almighty and all those persons who inculcated in me a desire to go ahead in the field of Library and Information Science. I would like to express my deep gratitude to all those who gave me the conviction to complete this thesis.

My guide Prof. J.N. Gautam, School of Studies in Library and Information Science, Jiwaji University, Gwalior, understood the potentials – and pitfalls – of Bibliometric studies. He placed confidence in my explorative and uncertain idea about the Zipf's law that I had while writing my dissertation at NISCAIR while doing the Associateship In Information Science. He gave me the decisive push to apply for a PhD scholarship in 2003. Having Prof. Gautam as a main supervisor on the PhD project has been a benefit. His enormous perception, originality and creative mind have been a huge force on this path. He has supported me to pursue my ideas and has been there every time I needed advice. His warm support, concern and confidence have been invaluable. Furthermore, I want to express my deep gratitude to Prof M.T.M. Khan, Prof & Head of Library & information science, Bundelkand University, Jhansi whose help, stimulating suggestions and encouragement helped me while doing research and writing this thesis. I am very grateful for the positive and constructive feedback given by him from time to time.

I would also like to express my gratitude to Dr. B.M. Gupta, NISTADS for his great and warm support. Many thanks are due to all my friends and batch mates at NISCAIR and St. John's College. They had provided an invaluable help and support during these years. I am thankful to Dr. Anurag Saxena, School of Management, IGNOU for his great kindness and humor, as well as his large professional and technical skills. I am also grateful to him for many inspirational and helpful discussions since my undergraduate years.

My parents Shri D.S. Saxena and Dr. Reeta Saxena have always encouraged me to follow my interests and I am thankful for the stimulation that they gave me in pursuing my wish to become a researcher and explore new frontiers in library and information science. I would also like to express my gratitude towards my father-in-law Dr. B.M. Jauhari and my brother Lt. Col. Anupam Saxena who have been extremely supportive of my endeavor. Last, but not least, I want to thank my husband Er Mohit Jauhari, without whose support and patience, I would never have accomplished this dissertation.

List of Tables

List of Figures

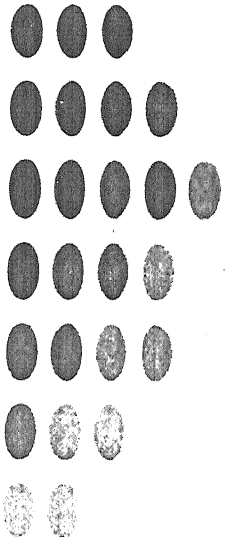
List of Abbreviations

Abbreviation	Meaning
AD	Anno Domini (Latin : "In the year of (Our) Lord")
ANSI	American National Standards Institute
ASCII	American Standard Code for Information Interchange
BC	Before Christ, also sometimes called BCE (Before the Common Era)
CDAC	Centre for Development of Advanced Computing
DGX	Discrete Gaussian Exponential
G-type	Gaussian type
HTML	Hypertext Markup Language
ICD9-CM	International Classification of Disease 9th Revision, Clinical Modification
IIT	Indian Institute of Technology
ITRANS	Indian Languages Transliteration
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy
KOSDAQ	Korean Securities Dealers Automated Quotations
KSE	Korean Stock Exchange
LIS	Library & Information Science
LLS	Linear least squares
LM algorithm	Levenberg-Marquardt algorithm
MIN	Minimum chi-square
MLE	Maximum likelihood
MOM	Method of moments

NHPP	Non-homogeneous Poisson process
PDF	Probability density function
PLIL	Pseudo Lingua for Indian Languages
QED	quod erat demonstrandum (meaning, "which was to be demonstrated").
RAT	Ratio of frequencies
SPSS	Statistical Package for the Social Sciences
US-ASCII	American Standard Code for Information Interchange
www	World Wide Web
Z-type	Zipfian type

Chapter 1

Introduction



Introduction

Zipf's Law

George Kingsley Zipf was born on Jan 7 1902 in Freeport, Ill. He graduated in 1924 from Harvard, summa cum laude. He studied German studies at Bonn in 1925. In 1929, he published a dissertation on "Relative Frequency as a determinant of phonetic change". He was awarded a Ph.D. in 1930 on comparative philology from Harvard. He taught German language as Assistant professor of German (Harvard) in 1936 and as University lecturer (Harvard) in 1939. He had interests in studying phonetic changes and thus worked on the frequency of "phonemes". His work was more philosophical rather than mathematical in nature. Many other researchers tried to find a mathematical foundation of his work. He died on Sep 25 1950 at the age of 48 only. Rousseau[1] (2002) presented a short biography Zipf and discussed his influence in the field of Informetrics and some recent applications of Zipf's law in Internet research, geography and economics.

As per Hertzel[2] (1987), Zipf had an idea that "speech as a natural phenomenon" is really "a series of communicative gestures" and after extensive research found that "the length of a word, far from being a random matter is closely related to the frequency of its usage-the greater the frequency, the shorter the word". Zipf also discovered that the "distribution of words in English approximates with remarkable precision a harmonic series... an unmistakable progression according to the inverse square, valid for well over 95% of all the different words in the sample".

Zipf formulated a law in 1930 that says frequency count (number of occurrence) of words in any text is inversely proportional to the rank of that word. In other words, the distribution of words adhered to a regular statistical pattern or "The probability of occurrence of words or other items starts high and tapers off exponentially. Thus, a few occur very often while many others occur rarely" (Black[3], 2000).

To further explain the basic form of the law,

frequency * *rank* has a inversely proportional relationship:

frequency * *rank* = *constant* or $f * r = c$ or $\log r + \log f = \log c$

Frequencies count of the words is the number of occurrences of the words in that text. The words are then arranged in the decreasing order of frequency so that the

most frequent word gets the highest rank. The frequency counts of words put in the same dictionary entries are regarded as the same. Zipf, in his first thesis, "Relative Frequency: A Determinant of Phonetic Change" wrote, "Observing the speech of many hundreds of millions of people, we have demonstrated, in part actually, in part by induction, that the conspicuousness or intensity of any element of language is inversely proportionate to its frequency. Using X for frequency, and Y for Conspicuousness (rank) we express our thesis thus: $Y = \frac{n}{X}$ *or* $XY = n$, where n is some constant, the actual size or value of which need not be our immediate concern now".

Zipf's Law approximates the relationship between rank and frequency of any text. The text should consist of at least 5000 words in order for the product of $r * f$ to be reasonably constant. Hřebíček & Luděk[4] (2002) discussed the questions related to corpus of more than 5000 words and discussed tautology in connection with the Zipf law.

Zipf attributed this law as a consequence of "Principle of Least Effort". The Principle of Least Effort postulates that a person would like to communicate in such a way as to minimize his total effort. According to Hertzel[2] (1987), "In simplest terms the Principle of Least Effort means, for example, that a person in solving his immediate problems will view these against the background of his probable future problems, as estimated by himself". In other words, a person will tend to "minimize" the probable average of his work-expenditure (over time), meaning use of least amount of work. Principle of Least Effort is relevant even today. If one have Internet access to resources, he is more likely to use it than the library. Altmann[5](2002) commented that Zipf's ideas are the foundation stones of modern quantitative linguistics and his influence is not restricted to linguistics but incessantly penetrates other sciences. According to Tague & Nicholls[6] (1987), "The Zipf's distribution plays a central role in the modeling of human activities, particularly of the variable studied in Bibliometrics and Scientometrics- productivity of researchers in a discipline, impact of authors or publications, use of words in a text or keys in a database and dispersion of a subject literature among sources". However Rapaport[7] (1957) commented "Zipf's arguments are vague appeals to the recognition of the principle in a great variety of situations simply on the basis of its

plausibility. And even these appeals are often stretched far beyond ordinary credibility".

Wyllys[8] (1981) made a special study of Zipf's law and called it "one of the most puzzling phenomena in bibliometrics". Wyllys[8] (1981) summarized the impact of the law as, "It is remarkable in its range of applicability to diverse phenomena, but we have not progressed far in an understanding why it should exist and why it should be so widespread". Wyllys[8] suggests that different slopes of Zipf's curves may characterize different subject fields. Another property of Zipf's law is that rank/frequency approximation is much better for the middle ranks than for the very highest ranks and the very lowest ranks.

Zipf [9] (1949) in his work, "Human Behavior and the principle of least effort" viewed language as a "tool" that is shaped by its "jobs" in human society. The purpose of this book, which was an introduction to human ecology, "is to establish the Principle of least effort as the primary principle that governs our entire individual and collective behaviour of all sorts". The study introduced the idea that behaviors that are "useful" are performed frequently, and frequent behaviors become quicker and easier to perform. The very existence of these quick, easy behavior patterns then causes individuals to choose them, even when they aren't necessarily the best behavior from a functional point of view. As per Zipf, "An investigator who undertakes to propound any such primary scientific principal of human behaviour must discharge three major obligations" that is, have a large verifiable numbers, be consistent, and have an understandable presentation". One observation of Zipf is "the greater the prestige of a person, the ever greater will be his power of attraction both for students and for grants of research money for the employment of technicians and for the purchase of expensive apparatus, with the result that his probable opportunities for making and reporting new 'important' and 'interesting' observations will tend to increase exponentially (i.e., 'nothing succeeds like success')".Other works of Zipf were "Selective Studies and the Principle of Relative Frequency in Language[10]" which as published in 1932, "Psycho-Biology of Languages[11]" which was published in 1935 and "National Unity and Disunity: The Nation as a Bio-Social Organism[12]" which was published in 1941.

In the study "Psycho-Biology of Languages" Zipf's goal was to put language study on a par with exact sciences, by use of "statistical techniques". It was an attempt to

prove that the key to the explanation of all synchronic and diachronic language-phenomena has been found in a statistically estimated tendency to maintain equilibrium between size and frequency. As per Hertzel[2] (1987), "Zipf recognized that there had been accurate investigative studies of language for about 100 years but nothing has ever been found in the nature of speech in any of its manifestations which is not completely comprised in the statement that speech is but a form of human behaviour".

According to Wyllys[8] (1981), "Zipf appears to this writer to have been poorly trained for dealing with quantitative phenomena. His knowledge of mathematics was minimal; of statistics, apparently nonexistent. He never showed interest in exploring the quantitative nature of his data beyond noting that they came close to his model of the moment. This done, he would launch into lengthy speculations about hazily defined possible causes. It is a pity that he almost never collaborated with statisticians. On the other hand, he was an indefatigable worker, and pursued the rank-frequency phenomenon and related ideas for twenty years despite often harsh criticism. There can be little doubt that the ubiquity of these phenomena would be less well recognized were it not for his work".

Madelbrot[13] (1953) tried to discuss Zipf's law in terms of communication costs and explained that the communication costs increases as the number of words and their length grows. Many years after his death linguistics agreed that speakers simplify communication by using a small pool of words that they can retrieve quickly from their memory and listeners simplify communication by preferring words with a single and unambiguous meaning. This proved that Zipf's law is applicable in understanding human language.

Definition and Mathematical Foundations of Zipf's Law

Chen and Leimkuhlar[14] (1986) stated that if one takes the words making up an extended body of text and ranks them by their number of occurances, then the rank r multiplied by its corresponding frequency of occurance, g(r) will be approximately constant, that is,

$g(r) = ar^{-1}, r = 1,2,3...$, where a is a positive constant.

There has been a debate as to if Zipf's law follows a Power-law or "stretched exponential" (Weibull) or "log-normal" or "Yule distribution".

More analysis showed that number of different words (N) of the same f-integral frequency of occurrence (under the conditions of the equation r x f = c) will be inversely proportional to the square of their frequency (approximately) or, stated some what more precisely in equation form, $N\left(f^2-\frac{1}{4}\right)=C$

This is Zipf's second law and has been called his "weak" law. Zipf's second law is also known as the discrete Pareto distribution[15] (1897), which involves count of vocabulary words (c_f) and their frequency. It states that

$$c_f \propto 1/f^{\theta}$$

Bi et al. [16] (2001) explained Zipf distribution and the two Zipf "laws": the rank-frequency one and the frequency-count one. The laws are best described with an example, such as words in a book (or the Bible, as we show in Figure 1) Let V be the vocabulary size, f_1 the occurrence frequency of the most frequent vocabulary word, and f_2 the second most frequent, and so on.

Definition 1: *The* rank-frequency *plot is the plot of the occurrence frequency f_r versus the rank r, in logarithmic scales*

The rank-frequency version of Zipf's law states that

$$f_r \propto 1/r$$

This is typically referred to as the *Zipf's law* or the *Zipf distribution.* In log-log scales, the Zipf distribution gives a straight line with slope -1.

The *generalized* Zipf distribution (or "Zipf-like" distribution) is defined as

$$f_r \propto 1/r^{\theta}$$

where the log-log plot can be linear with any slope.

The second 'law', also known as the discrete Pareto distribution, involves the \count-frequency" plot: let c_f be the count of vocabulary words that appear f times in the document. The second Zipf's law states that

$$c_f \propto 1/f^{\theta}$$

Many scientists have analyzed, refined and evaluated Zipf's endeavors, but Wyllys[8] who has made a special study of Zipf's law, called it "One of the most puzzling phenomena in Bibliometrics" and noted that the Zipf's law only approximates the relationship between rank r and frequency f for any actual corpus. Zipf's work showed that the approximation is much better for the middle ranks than that for the very lowest and the very highest ranks, and his work with samples of various sizes suggest that the corpus should consist of at least 5000 words in order for the product r X f to be reaosnably constant, even in the middle ranks.

Tague & Nicholls[6] (1987) commented that in general, Zipf's law may be described as representing the distributions of a set of tokens over a set of types. It has been represented in a number of functional forms, which may be distinguished by the number of parameters and by the nature of property or variable described, whether a size (frequency) or a rank. Zipf's distribution resembles in structure to many other distributions such as the Yule and Bradford distributions, and Lotka's law. Each one of them has an empirical regularity in the study of many diverse subjects.There are four major school of thought on the theoretical underpinning of Zipf's law. The following table demonstrates them.

Major thoughts on the Zipf's Law

Person	Direction of Thought
Φ Hill & Woodroofe Sichel, Crowley, Bliss	Zipf's law can be derived from stochastic processes [Hill[17] (1970)a, b[18], Hill & Woodroofe [19] (1975), Sichel[20] (1975), Crowley[21] (1975), Bliss[22] (1953)]
Φ Hill	Bose – Einstein form of the classical occupancy model
Φ Bliss/Fisher	Negative Binomial Model
Φ Simon/Price	Many of the classical occupancy model can be manipulated to yeild hyperbolic distributions. Simon[23](1960), Price[24] (1976)
Φ Simon	Beta function
Φ Price	Cumulative advantaged distribution
Φ Mandelbrot	Information theoretic approach to study the statistical structure [Mandelbrot[13] (1953)].
Φ Herdan	Works based on the field of quantitative linguistics [Herdan[25] (1964)]
Φ Haitun Brookes	Laplace's law of succession is shown to be the 'Zipfian' frrequency analogue of the Bradford Law. [Brookes[26] (1984)/ Haitun[27] (1982)]

Table 1.1: *Researchers & their direction of thought about Zipf's law*

Let us see, the works of different persons to get an insight.

Hill's Derivation

Hill[17] (1970) derived Zipf's Law from Bose-Einstein form of classical occupancy model with a random number of cells. It was proved that an extension of the Bose-Einstein model of allocations within regions yeilds convergence to a form of Zipf's law. (Generic specific form). It is described as a system of classification of units such that the proportion of classes with exactly 's' units is in some specified sense approximately proportional to $s^{-(1+\alpha)}$ for some $\alpha > 0$, with $\alpha \leq 1$ as a case of interest (Hill & Woodroofe[19], 1975). This eventually is equivalent to Zipf's first law.

The model proposed by hill cn be described as follows. Suppose there are N species which are to be distributed to M nonempty genera. Let L_i be the number of species allocated to genus I, and let G(s) be the resulting number of genera with exactly s species. Suppose that the allocation of species to genera is of Bose-Einsteen form

$$\Pr\{\underline{L} = \underline{l} | M, N\} = \binom{N-1}{M-1}^{-1}$$

for all $\underline{l} = (l_1, \ldots l_M)$ such that $l_i \geq 1$, $\sum_{i=1}^{M} l_i = N$

Suppose further that given N, M has a conditional distribution such that $\Pr\{M / N \leq x \mid N\}$ converges properly to a distribution F(x) with F(0) = 0. Then it was shown that G(s) / M, the proportion of genera with s species is in limits as $N \rightarrow \infty$, distributed like $\Theta (1-\Theta)^{s-1}$ where Θ denotes a random variable having distribution F. If Θ has a beta distribution B(a,b), i.e., if F has density function

$$F'(x) = \Gamma(a+b)[\,\Gamma(a)\Gamma(b)\,]^{-1} x^{a-1}(1-x)^{b-1},$$
where Γ *is the Gamma function,* $0 \leq x \leq 1$, *and* $a > 0, b > 0$

Then,

$$E\{\Theta(1-\Theta)^{s-1}\} \approx a\Gamma(a+b)[\Gamma(b)]^{-1} s^{-1(1+a)}$$

As $s \rightarrow \infty$, where the symbol "~" indicates that the ratio of the two sides tends to unity. In fact this approximation is generally good even for small s. For example, if Θ has the uniform distribution on the unit intervel, then

$$E\{\Theta(1-\Theta)^{s-1}\} = [s(s+1)]^{-1}$$

This is a simple and important form of Zipf's law fitting approximately a great variety of data. Thus they presented a conceptually simple model for an exceedingly complex phenomenon. The most fundamental assumption, that underlying all the theory is the approximate Bose-Einstein allocation of species to genera within a family.

Price's Derivation

Price[24] (1976) postulated that it is based on cumulative advantage distribution, which can be derived from a modification of the Polya Urn model, or as a stochastic birth process.

Let us consider a population of n_T individuals, of whom, a fraction $f(r)$ are in state r, where r is the total of "success" (occurances) thus far achieved by each of the individuals in the fraction $f(r)$ of the population,

$$\sum_{1}^{\infty} f(r) = 1$$

& the mean number of previous "successes"

$$\sum_{1}^{\infty} r\, f(r) = R$$

If there are further 'successes" individual will move from state r to r+1.

Now suppose a small number dn_T of new individuals are added to the population, under with them Rdn_T new successes are sprinkled evenly at random over all members, there will be dn_T/n_T new successes per previous ones, and for the class of $n_T f(r)$ individuals with r previous successes each, there will then be $rn_T f(r)/n_T$ new successes, and therefore transitions from this rth state to be (r+1)th state, leads to $rf(r)\, dn_T$ transitions into it from the class below receiving its quota of new successes. The change in the number of individuals in the rth state is therefore

$$n_T \frac{d}{dn_T} f(r) = \begin{cases} -rf(r)+(r-1)f(r-1), & ,r>1 \\ -f(1) & ,r=1, \end{cases}$$

so that

$$n_T \frac{d}{dn_T} f(r) = \begin{cases} -(r-1)f(r)+(r-1)f(r-1) & ,r>1 \\ -2f(1)+1 & ,r=1, \end{cases}$$

and the distribution over the states is defined by this series of difference-differential equations. For a stable distribution for which f(r) is independent of n_T

$$\frac{d f(r)}{dn_t} = 0$$

$$\Rightarrow f(r) = \frac{r-1}{r+1} f(r-1)$$

$$= \frac{r-1}{r+1} \frac{r-2}{r} \frac{r-3}{r-1} \cdots\cdots \frac{1}{3} \frac{1}{2}$$

$$or\ f(r) = \frac{1}{r(r+1)}$$

Which is the form for the urn model.

Mandelbrot's Derivation

Mandelbrot[13] had published several studies of generalizations of Zipf's law dealing with the question of whether the slope is -1 and with the deeper problem of explaining why the *rf* products should be relatively constant. Mandelbrot (1952, 1964) assumed that the aim of language is to transmit the most information per symbol with the least effort. Following relationship is obtained.

$$f(r) = K(r+c)^{-\theta}$$

Where,

f(r) is the rank frequency and r is the rank of the word and c & θ are constants, c improves the fit for small r and the exponent θ improves the fit for large r.

Mandelbrot showed that previous equation is similar to a regular lexicographical tree. He defines a lexicographical tree as one having (N+1) trunks, numbered 0 through N, where the first trunk corresponds to a space (Empty word) and each of the others corresponds to a letter. Each of the "Letters" trunks has N+1 branches corresponding to the space and N letters. The space branch is again barren, and the others branch N+1 times each, and so on. The end of each branch corresponds to a word with a given probability [f(r)]. Booth[28] (1967) proved that Mandelbrot's derivation and Zipf's Law are equivalent.

Herdan Waring Derivation

Herdan[25] (1964) described Language as a coding system composed of individual speech utterances and the different words found in Language, while emphasising the independence of sound and meaning. Herdan then presented the following model of vocabulary frequency whose starting point is Waring's expansion for

$$\frac{1}{p-q}, \quad \textit{i.e,} \text{ Where } p>0,\ q>0$$

$$\frac{1}{p-q} = \frac{1}{p} + \frac{q}{p(p+1)} + \frac{q(q+1)}{p(p+1)(p+2)} = ------ + \frac{q[r]}{p[r+1]} + ----$$

Multiplying both sides by p-q

$$1 = (p-q)\left(\frac{1}{p} + \frac{q}{p(p+1)} + ---- + \frac{q(q+1)}{p(p+1)(p+2)} + ----- + \frac{q[r]}{p[r+1]} ----\right),$$

Where the r^{th} term represents $f(r)$ in the frequency distribution

The mean & variance of the distribution are,

$$\mu = \frac{q}{p-q-1} \quad \& \quad \sigma^2 = \frac{q(p-1)(p-q)}{(p-q-1)^2(p-q-2)}$$

Where

$$f(r) = \frac{(p-q)(q)(q+1)----(q+r+1)}{p(p+1)(p+2)---(p+r+2)}$$

For r = 2, 3, ..., p & q are such o<q<p & $f(r)$ is the probability that a word will appear with frequency r in large text.

One can see that Zipf distribution can be derived from a beta function. It was suggested to have a theoretical justification for applying his derivation to text word distributions. The model is based on the fact that the author choose words according to the process of imagination, association and imitation (Fedorowicz[29], 1982)

Haitun[27]/ Brookes[26] Derivation

Zipf, the professional linguist was more interested in his own field rather than statistics. But he accepted the statistical regularities found in his work. In order to do this Zipf deviated from the frequency distributions of orthodox statistics and postulated frequency rank distributions. In learning the vocabulary of a new

language, the words of most immediate interest are those which occur most frequently in texts in that language. There is a rough and ready rule of the thumb- the 80/20 rule-which states that 80% of the bibliography or of the language text is provided by the most productive 20% of the sources. Zipf adopted the unorthodox statistical technique of ranking their sources, beginning with the most frequent. The advantage of ranking is that it brings to the forefront of the distribution those items of greatest interest and relegates to the distant tail those items of rare occurrences which are relatively difficult to find and identify- thus reversing the procedure imposed by frequency distribution.

As Zipf's law is concerned with 'categorization', let us see whether the Laplace law is related to them;

The number of items, and therefore the number of entities to be ranked, in the tail of the Laplace distribution from $x=m$ to its end point at $x=n+1$ is given by

$$r = \frac{k}{m} - \frac{k}{n+1}$$

As both k and $(n+1)$ are constants, we can put $\frac{k}{n+1} = w$ and rewrite the relation as

$$\frac{k}{m} = r + w$$

The number of items embraced by the Laplace law over this same range, m *to* $(n+1)$ is given by

$$G(r) = \int_{m}^{n+1} \frac{k}{x^2} . x\, dx = k \log_e (n+1) - k \log_e m$$
$$= k \log_e (k/w) - k \log_e k/(r+w),$$
$$= k \log_e (1 - r/w)$$

This equation is formally identical to the formulation of Bradford Law. As Zipf did not cumulate the frequency of his f/r data, the Zipf law is given by

$$g(r) = \frac{dG(r)}{dr} = k/(r+w)$$

This is one of the forms proposed by Zipf.

The purpose of the series of derivations is to demonstrate the process by which the Zipfian approach to word distribution modelling can be represented, and also to establish a mathematical basis for its use in bibliographic database retrieval systems. Zipf's law is described in terms of the underlying processes governing the choice of texts words (and, indeed, many other phenomena). These processes also direct the distribution of the contents of the inverted file of such a database system, since the inverted file is merely an alternate method of arranging the words contained in the textual material (Fedorowicz[29], 1982).

Some more approaches to Zipf's Law

Bi et al.[16] (2001): *Zipf's law as a special case of Discrete Gaussian Exponential (DGX) distribution*

Bi et al.[16] (2001) presented Zipf's law as a special case of DGX distribution. Their goal was to find a discrete distribution that will fit the PDF (a.k.a frequency-count plot) of many, real data sets. There were many options to fit distributions like parabola, third degree polynomial, gaussian, sinusoid and splines etc. But question arises: even if one of these functions fits in a few cases, do one has "a-priori" reasons to believe that it will fit well, in multiple settings?

According to them, the answer to all this questions is proposed DGX distribution. Judging from the success of the lognormal (also referred to as "anti-lognormal") distribution for continuous data, they proposed the following thought experiment:

Consider a random variable, say, the duration of a web-surfing session. This is a continuous variable, and, most likely, might follow a lognormal distribution. However, we need to store it with finite accuracy, and thus turn it into an integer (number of minutes, or seconds, or hours). This is exactly the motivation behind DGX.

Consider a lognormal random variable (by creating a Gaussian variable and exponentiating it); then, digitize it to the nearest integer. The same is true for everything else: salaries (digitized to penny accuracy), duration of hospital stays (rounded to days), body height (inches), body weight (pounds) and so on. There is a subtle, but important point: If the lognormal random variable becomes zero after the rounding, we omit it. This is necessary, since, e.g., we don't know how many vocabulary words have not appeared in our document. Notice that this omission

leads to the so-called "truncated" or "veiled" random variables, which are notoriously difficult with respect to their parameter estimations, in the continuous case.

They presented their proposed discrete PDF. They proposed a distribution with the following PDF:

$$P(x=k)=\frac{A(\mu,\sigma)}{k}\exp\left[-\frac{(\ln k-\mu)^2}{2\sigma^2}\right] \quad k=1,2,...$$

Where

$$A(\mu,\sigma)=\left\{\sum_{k=1}^{\infty}\frac{1}{k}\exp\left[-\frac{(\ln k-\mu)^2}{2\sigma^2}\right]\right\}^{-1}$$

is a normalization constant depending on μ and σ. This PDF has the following characteristics

- It is discrete, which means it is suitable to model many real discrete distributions.
- It is a discretized version of a known continuous distribution, the lognormal distribution. As we know, the PDF of a lognormal distribution is a parabola in log-log plot, which is next simplest model beyond a straight line.
- This model has only two parameters to estimate, so it is not difficult to compute.

Zipf's law as a special case

Lemma 1. *The Discrete Gaussian Exponential (DGX) as defined by proposed PDF reduces to Zipf's law as* $\mu\rightarrow\infty$

Proof: We first rewrite proposed PDF as

$$P(x=k)\propto\frac{1}{k}\exp\left(-\frac{\ln k(\ln k-2\mu)}{2\sigma^2}\right)$$

Assume that $\ln k=|\mu|$, the PDF becomes

$$P(x=k)\propto\frac{1}{k}\exp\left(-\frac{\mu\ln k}{\sigma^2}\right)\propto k^{-1+\mu/\sigma^2}$$

which reduces to generalized Zipf distribution with slope $\varphi = 1-\mu/\sigma^2$. QED

DGX works well on real data sets both when their PDF has a clear curvature and when the PDF is straight in log-log plot.

Parunak Anita[30] (1979): *Formula that relates word length to the rank-size*

They developed a graphical means of comparing sets of ordered counts (of different overall size), without assuming the form of the distribution. Such a technique once available can also be used to display goodness of fit of specific analytic forms. In using the technique to compare different sets of word-frequency data, a formula was discovered that related word length to the rank-size rule.

Suppose we have a total of 'T' items distributed over 'D' cells. The cell frequency (or size or count, F_i, is the number of items in cell i. Thus,

$$\sum_{i=1}^{D} F_i = T$$

When number ordered from greatest to least, the F_i may be denoted by

$$F_{(1)} \geq F_{(2)} \geq \ldots \geq F_{(r)} \geq \ldots \geq F_{(D)}$$

The approximate relation

$$r\,F_{(r)}^{\rho} \approx d, \quad r = 1,2,\ldots,D$$

where d and ρ (≥ 1) are constants

This relation is called the rank-size rule or Zipf's Law. Notice that if some ranks are tied (that is if $F_{(r)} = F_{(r+1)} = \ldots = F_{(r+a)}$ for some integer a), then the products $rF_{(r)}, (r+1)F_{(r+1)}, \ldots, (r+a)F_{(r+a)}$ cannot be equal. Also a plot of $F_{(r)}$ against r is J-shaped and usually very long tailed, because of the existence of many tied ranks, especially at low frequencies $(F_{(D)}, F_{(D-1)}, \ldots)$

Tague & Nicholls[6] (1987): *Zipf's function describes the distribution of a set of 'm' tokens over a set of 't' types*

In its most general form, the Zipf's function describes the distribution of a set of 'm' tokens over a set of 't' types using one of the following expressions:

$$g(x) = \frac{a}{(x+c)^b}, \quad x = 1,2,\ldots,x_{max}, \quad a,b > 0, c \geq 0$$

$$f(r) = \frac{a'}{(r+c')^{b'}}, \quad r = 1,2,....,t, \; a',b' > 0, c' \geq 0$$

Where g(x) is the number of types with exactly x tokens and f(r) is the number of tokens for the r^{th} ranking type when types are arranged in descending order of number of tokens. The function g(x) is commonly called a size-frequency distribution, as opposed to f(r), which is a rank-frequency distribution

The parameter x_{max} represents the maximum number of tokens for a type, or the maximal size or value of the productivity variable x. Note that x_{max}=f(1), that is the frequency of the highest ranked type. In most assumptions, c' is assumed to be 0, that is

$$g(x) = \frac{a}{x^b}, \quad x = 1,2...,x_{max}, \quad a,b > 0$$

In this case, the parameter a will represent the number of types with exactly one token. The larger the exponent b, the larger will be this number relative to the total number of types.

The Zipf's size frequency distribution cn be expressed as a relative frequency or probability distribution by dividing by a suitable constraint. If X represents the number of tokens assigned to a random type, p(x) the probability X assumes a specific value x, and t the total number of types, then

$$p(x) = \frac{g(x)}{t} = \frac{a}{tx^b}, \quad x = 1,2....,x_{max}, \quad a,b > 0$$

The size variable X can be generalized to a continuous productivity variable, that is, productivity of a type rather than number of tokens of a type. The discrete Zipf distribution is then replaced by its continuous analogue, the Pareto distribution.

Lee Breslau et al.[31] (1999): *Zipf-like distribution with the varying exponent*

Consider a cache that receives a stream of requests for web pages. Let N be the total number of web pages in the universe. Let $P_N(i)$ be the conditional probability that, given the arrival of a page request, the arriving request is made for page i. Let all the pages be ranked in order of their popularity where page i is the i^{th} most popular page. We assume that $P_N(i)$, defined for i = 1,2,....N, has a "cut-off" Zipf-like distribution given by

$$P_N(i) = \frac{\Omega}{i^{\alpha}},$$

Where

$$\Omega = \left(\sum_{i=1}^{N} \frac{1}{i^{\alpha}} \right)^{-1}$$

The true Zipf's law has $\alpha=1$ but if one considers a broader class of distribution functions with exponents in the range $0<\alpha\leq1$, each page request is drawn independently from the Zipf's distribution.

References

1. Rousseau, Ronald. (2002). George Kingsley Zipf: life, ideas, his law and Informetrics. *Glottometrics 3*, pp 11-18, 2002.
2. Hertzel, Dorothy. H. (1987). History of the development of ideas in bibliometrics. *Encyclopedia of Library & Information Sciences,* Vol. 42, Supplement (7), pp 180-219.
3. Black, Paul. E. (2000). Zipf's Law: Definition, at http://hissa.nist.gov/dads/HTML/zipfslaw.html. Site accessed on 1/29/01.
4. Hřebíček, Luděk. (2002). Zipf's law and text, *Glottometrics 3*, pp 27-38, 2002
5. Altmann, Gabriel. (2002). Zipfian linguistics. *Glottometrics 3*, 9-26.
6. Tague, Jean. & Nicholls, Paul. (1987). The maximal value of a Zipf size variable: Sampling properties and relationship to other parameters. *Information Processing & Management,* Vol. 23, No.2, pp 155-170.
7. Rapaport, Anatole. (1957). The Stochastic and the 'Teleological' Rationales of Certain Distributions and the So-called Principle of Least Effort, *Behav. Sci.*, 2, 150.
8. Wyllys, Ronald. E. (1981). Empirical and Theoretical Bases of Zipf's Law. *Library Trends* 30(1) (Summer 1981): 53-64
9. Zipf, G. K. (1949). *Human Behavior and the principle of least effort.* Cambridge, MA: Addison-Wesley Press
10. Zipf, G. K. (1932). *Selective Studies and the Principle of Relative Frequency in Language.*
11. Zipf, G.K. (1935). *Psychobiology of Languages*, Houghton-Mifflin, 1935; MIT Press.
12. Zipf, G.K. (1941). *National Unity and Disunity: The Nation as a Bio-Social Organism*, Principia Press, Bloomington Indiana, 1941.
13. Mandelbrot, B. B. (1953). An informational theory of the statistical structure of languages, in *Communication Theory*, ed. W. Jackson (Butterworth, 1953), pp. 486-502.

14. Chen, Ye-Sho. & Leimkuhler, Ferdinand. F. (1987). Analysis of Zipf's Law: An index approach. *Information Processing and Management*, 23(3): 171-182.

15. Pareto, V. (1897). *Cours d'Economie Politique*. Rouge and Cie, Lausanne and Paris.

16. Bi, Zhiqiang., Faloutsos, C. & Korn, F. (2001). The DGX distribution for mining massive, skewed data. *Conference on Knowledge Discovery and Data Mining (KDD)* 2001.

17. Hill, Bruce. M. (1970)[a]. Zipf's law and prior distributions for the composition of a population. *Journal of the American Statistical Association*, 65:1220-1232.

18. Hill, B. M. (1970)[b]. Rank Frequency form of Zipf's Law. *Journal of the American Statistical Association*. 69 (348): 1017-1026; 1974.

19. Hill, B. M. & Woodroofe, M. (1975). Stronger Firms of Zipf's Law. *Journal of American Statistical Association*. 70 (349); 212-219; 1975.

20. Sichel, H. S. (1975). On a distribution Law for word frequencies. *Journal of the American Statistical Association*. 70 (352) part I 542-547; 1975.

21. Crowley, C. J. (1975). The Distribution of Citation to Scientific Papers: A Model, presented at *Midwest Sociological Society Meeting*, Chicago, April 1975 (unpublished).

22. Bliss, C. I. & Fisher, R. A. (1953). Fitting the Negative Binomial Distribution to Biological Data & Note on the Efficient Fitting of the Negative Binomial. *Biometrics* (2): 176-200; 1953.

23. Simon, H. A. (1960). Some further notes on a class of skew distribution functions. *Information and Control*, 3, 80-88.

24. Price, Derek. deSolla. (1976). A General theory of Bibliometric & other Cumulative Advantage Processes. *Journal of the American Society for Information Science*. 27(5); 292-306; Sept-Oct 1976.

25. Herdan, G. (1964). *Quantitative Linguistics*. Washington, D.C : Butterworths, 1964.

26. Brookes, B. C. (1984). Towards Informetrics: Haitun, Laplace, Zipf, Bradford and the Alvey programme. *Journal of Documentation*, Vol. 40, no. 2, pp120-143.

27. Haitun, S. D. (1982). Stationary Scientometric Distributions. Part I. The different approximations. *Scientometrics*, 4(1), 5-25. Part II. Non-Gaussian nature of scientific activities, *Scientometrics*, 4(2), 89-104.Part III. The role of Zipf distribution, *Scientometrics*, 4(3), 181-94.

28. Booth, A. D. (1967). A law of occurrences for of low frequency. *Information & Control*. 10(4): 386-393; April, 1967.

29. Fedorowicz, Jane. (1982). The Theoretical foundation of Zipf's Law and its application to the bibliographic database environment. *Journal of the American Society for Information Science*, pp. 285-293, Sep. 1982.

30. Parunak, Anita. (1979). Graphical analysis of ranked counts (of words). *Journal of the American Statistical Association,* Volume 74, No- 365.

31. Lee, Breslau., Pei, Cao., Li, Fan., Graham, Phillips. & Scott, Shenkar. (1999). Web Caching and Zipf-Like Distributions: Evidence and Implications, *IEEE INFOCOM*, Vol. XX. NO- Y.

Chapter 2

Review of Literature

There are many more attempts in proving that Zipf's law is actually a Power-law or "stretched exponential" (Weibull) or "log-normal" or "Yule distribution". To mention a few, Yule distribution (Martindale[1] et al., 1996), Log-normal distribution (Perline[2], 1996) and Stretched exponential distribution (Laherrere[3] et.al,, 1998).and double-pareto lognormal distribution (Reed[4-8], 2001, 2002, 2003). Bi[9] et al. (2001) proposed an alternate distribution called DGX, which included Zipf and generalized Zipf distributions as special cases. They commented, "...the Zipf distribution often fails to model real data sets well... the Zipf (or generalized-Zipf) distribution would expect the plots to be straight lines in logarithmic-logarithmic scales. However, we observe a clear tilting. Zipf himself had observed this deviation and even had a name for it (top concavity"), and he devoted several paragraphs in his book to justify it, whenever it appeared in a data set".

Laherrere[3] et al. (1998) commented that "Power laws are generally used to represent natural distributions, often claimed to be power laws which represent as linear regressions in log-log plots. In reality however, the plots often display linearity over a limited range of scales and/or exhibit noticeable curvature". They found that stretched exponential distributions provide a reasonable fit to all data sets and has the advantage of a sound theoretical foundation. Stretched exponentials also have the advantage of being economical in their number of adjustable parameters.

Mandelbrot[10] (1959) criticized Simon's model[11] (1955) concerning the class of frequency distributions generally associated with the name of G.K. Zipf. He commented that Simon's model is analytically circular in some cases. Simon[12] (1960) refuted this and commented that the basic parameter of the distributions is almost always very close to unity and hence simple stochastic models can be constructed. Mandelbrot[13] (1961) maintained his objections to Simon's 1955 model for the Pareto-Yule-Zipf distribution.

As cited in Hertzel[14] (1987), Wyllys[15] (1975) summated as "Inclined towards mysticism. Zipf not only leaped to the conclusion that the 'true' slope of rank-frequency curves was -1, but also claimed that this regular slope resulted from some

fundamental force of nature. In the broad sense, this claim had to be correct; but Zipf vigorously described the force as that of struggle between the 'Life Tendency' and the 'Death Tendency' or the 'Force of Diversification' and the 'Force of Unification' and finally as the 'Principle of Least Effort' for none of which did he furnish an operable definition. However, in work summarized...Zipf did show that an astonishingly wide range of phenomena...exhibited distributional behaviour that could be approximated by his 'Law'". As per Haitun[16] (1982), "Zipf's law applies to the distribution of many social characteristics, and that this law implies that social phenomena are inherently non-gaussian". Chai Kim[17] (1982) investigated the extent to which the principle of least effort as advanced by Zipf provided a theoretical basis for identifying and updating descriptors of science/technology and social sciences. He found that "the relative frequency of occurrence of the descriptors of social sciences conformed to the theoretical distribution of Zipf while that of science/technology did not".

According to Thom & Zobel[18] (1992), "Zipf's law is perhaps the best known model of word probabilities. It describes the fact that when words are ranked on frequency, from most to least frequent, plotting rank against frequency yields a hyperbolic curve". They argued that too much of emphasis has been placed on this result (Zipf's law).

Witten & Bell[19] (1990) found that even words produced by a simple random generator conform to Zipf's law. According to them, "Although theoretically elegant, Zipf's law provides only a loose fit to actual text and in practice must be modified by introduction of additional parameters".

The Russian statistician S.D. Haitun[16] published a three-part comprehensive review of all the empirical frequency distributions that have been reported in the literature of bibliometrics and related fields. He postulated that all the empirical distributions can be divided into two types. These are Gaussian type (G-type) and Zipfian type (Z-type). According to him G-type are those distributions that are characterized by the fact that these have as many higher moments as modern statistical theory demands. Z-type distributions have no moments whatever. According to Brookes (1984), "Gaussian-type distributions arise only in physical contexts; Zipfian only in social contexts. As the whole of modern statistical theory is based on Gaussian distributions, Haitun thus shows that its application to social statistics, including

cognitive statistics, is 'inadmissible'. A new theory based on Zipfian distributions is therefore needed for the social sciences".

Ivancheva[21] (2001) attempted to answer the question, "Why do most bibliometrics and scientometric laws reveal characters of Non-Gaussian distributions, i.e., have unduly long 'tails'. Ivancheva postulated a corollary that for a discrete pareto distributed random variable, $\alpha=1$ is the most reasonable value for family of Zipf laws, applied to information or social phenomena. Nicholls[22] (1987) applied many methods for estimation of Zipf parameters. These included linear least squares (LLS), maximum likelihood (MLE), ratio of frequencies (RAT), minimum chi-square (MIN), method of moments (MOM) and a truncated least squares method. Apostolos & Li[23] (1997) proposed a novel metric for the evaluation of the goodness-of-fit criterion between the distribution functions of two samples. They extended the usage of the proposed criterion for the case of the generalized Zipf distribution. According to them, "Since the Zipf distribution of a document employs the frequencies of the words forming that particular document; it is justified to evaluate the contextual similarity based on the numerical encoding produced by the particular distribution". Egghe[24] (1999) studied the probabilities of the occurrence of multi-word (m-word) phrases (m=2, 3 ...) in relation to the probabilities of occurrences of the single words. They found that in the latter case, the law of Zipf is valid.

Applications of Zipf's Law

Many researchers have applied Zipf's law in city populations. Hill[25] (1970) applied Zipf's law for the composition of a population. He found that limiting distribution of frequencies as the population size become large; the limiting distribution gets a weak form of Zipf's law. Makse[26] et al. (1995) modeled urban growth patterns and used Zipf's law. Krugman[27] (1996) used Zipf's law for the Self-Organizing Economy. Zanette[28] et al. (1997) developed a model of a large-scale city formation. Manrubia[29] et al. (1998), developed an intermittency model for urban development Marsili & Zhang[30] (1998) modeled interacting individuals and commented, "In many disparate societies, it is not unnatural to assume that individuals make their city-dwelling decision based on their own opinions as well as on their interaction with other citizens". They found that the larger cities obey approximately Zipf's law. Gabaix[31] (1999) gave an explanation for Zipf's law for cities. Reed[32] (2002)

analyzed the rank-size distribution for human settlements on the basis of simple stochastic models and found that model explains the rank size phenomenon in the upper tail. According to Reed[33] (2001), "It has long been recognized that the distribution of size (human population) of cities within a particular country or jurisdiction frequently exhibits Paretian behaviour in the upper tail. This phenomenon is known as the rank size property or in the case when the Pareto exponent is unity as Zipf's law. There have been many attempts to explain this phenomenon". Knudsen[34] (2001) found that the growth pattern of Danish production companies follows a clean rank-size distribution consistent with Zipf's law. They tested the existence of Zipf's law on 14, 541 Danish production companies and found answers to three basic questions like does the Danish case refute Zipf's law for cities, what arethe implications of Zipf's law for models of local growth? And do we have a Zipf's law for firms? Based on empirical data they found that the growth pattern of Danish production companies follows a clean rank-size distribution consistent with Zipf's law. Marsili and Zhang[30] (1998) presented a general approach to explain the Zipf's law of city distribution. They commented, "If the simplest interaction (pair wise) is assumed, individuals tend to form cities in agreement with the well-known Statistics". According to them, the interaction leading to Zipf's law is, on one hand, the simplest possible (pair wise interaction). On the other it is a rather special one, since it is the "lowest order" of interaction which does not lead to the formation of a mega city, which draws a good portion of the whole population. Urzua[35] (2000) presented a simple and locally optimal test for Zipf's law and illustrated its use in the case of the largest US metropolitan areas. He commented, "the log of the Zipf variate x/μ follows an exponential distribution with mean equal to one, while its inverse follows a uniform with mean equal to one-half.

Many scientists have attempted to examine the informetric properties of the web in the past. Adamic & Huberman[36] (2002) claimed the Zipf's law governs many features of the Internet. It has implications for the design and function of the Internet. The connectivity of Internet routers influences the robustness of the network while the distribution in the number of email contacts affects the spread of email viruses. Even web caching strategies are formulated to account for a Zipf distribution in the number of requests for web pages. According to Adamic & Huberman[36] (2002), the Internet is comprised of networks on many levels, and some

of the most exciting consequences of Zipf's law have been discovered in this area. The distribution of, the number of computers a computer has connections to, is a Zipf distribution. The presence of Zipf's law has implications for the search strategies used in P2P networks. Knowledge of Zipf's law in the connectivity distribution has offered a solution to an Internet communication problem. According to Shi[37] et al. (2006), "Zipf's law (Zipf-like law) holds the promise of more effective design and use of Web cache resources. Ongoing work includes the application of the work studied in this paper and the study of the Web prefetching model based on the Zipf's law". Rousseau[38] (2001) has tried to analyze a time series of the number of hits of word "Euro" on the web during a period of one year. Lee Breslau[39] et al. (1999) raised an issue that whether web requests from a fixed user community are distributed according to Zipf's law. They found that the page request distribution seen by web proxy caching using traces from a variety of sources does not follow Zipf's distribution precisely, but instead follows a Zipf-like distribution with varyirg exponents. Chao & D'haeseleer[40] (2001) attempted to find the distribution of Variable length Phatic interjectives on the World Wide Web. They found that the number of pages found containing these words would fall off as a power law. However the exponents for length frequency distributions of different interjectives were much larger than –1 predicted by Zipf's law. There are many other instances of Zipf's law in Web Access Statistics and Internet Traffic like caching relay for the world wide web (Glassman[42], 1994), Internet web server (Arlitt[41] et al. 1997), World Wide Web traffic (Crovella[43] et al., 1997), power laws in designed systems like Internet traffic (Carlson[44], 2000) and nature of markets in the World Wide Web (Adamic[36] et al. 2000). According to Chen & Wu[45] (1997), "Many models have been developed to predict a software system's failure rate and were used as management tools to evaluate software reliability. Software failure processes can be modeled by non-homogeneous Poisson process (NHPP), which was originally used to analyze the hardware failure data. The Duane model is a well-known NHPP model based on power law failure rate for analyzing hardware reliability". They proposed a model is proposed based on Zipf's law for software reliability analysis and observed that the proposed model has better long-term predictability than the Duane model for failure data sets with power law's failure rates.

Zipf's law has applications in finance and business also. If the distribution is not plotted as the rank-frequency plot, but the number of companies in each revenue or sales then Zipf's is observed. Champernowne[46] (1953) presented a model of income distribution. Mandelbrot[47] (1963) discussed new methods in statistical economics and also in his book *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk* (Mandelbrot[48], 1997). Aoyama[49] et al. (2000) applied Pareto's law for income of individuals and debt of bankrupt companies. Reed[50] (2001) commented that many empirical size distributions in economics and elsewhere exhibit power-law behavior in the upper tail. The prime examples are distribution of incomes (Pareto's law) and city sizes (Zipf's law on rank-size property). Stanley[51] et al. (1995) related Zipf's plots and the size distribution of firms. Axtell[52] (2001) also showed that the Zipf distribution characterizes firm sizes; the probability a firm is larger than size s is inversely proportional to s. According to them, The Zipf distribution is an unambiguous target that any empirically accurate theory of the firm must hit. This result places important limits on models of firm dynamics. Because the Zipf distribution obtains all the way down to the smallest sizes, it should be possible to derive Kesten-type processes and, hence, the Zipf distribution forms a microeconomic model in which individual agents interact to form productive teams. Bence & Oppenheim[53] (2004) tested the group of 1489 titles for a Bradford-Zipf distribution and posed a question whether Bradford-Zipf apply to business and management journals in the 2001 Research Assessment Exercise? According to them, "Zipf's Law describes the frequency distribution of words in a given text, with familiar words being used many times and many words being used only once. Bradford's and Zipf's laws have been shown to be mathematically identical and so the distribution is often referred to as the Bradford-Zipf distribution". Choi[54] et al. (2005) investigated the rank distribution and the cumulative probability for stock prices, and the probability density of price returns for stocks traded the Korean Stock Exchange (KSE) and the Korean Securities Dealers Automated Quotations (KOSDAQ) market. According to Choi[54] et al. (2005), "the ranks for stock prices traded on the KSE, the KOSDAQ, and the TSE follow Zipf's law or a power law while that of the NYSE follows a power law". As per Samuelsson[55] (1996), Zipf's Law is also closely related to the Good-Turing smoothing technique, and a better law could lead to better smoothing. He showed that Zipf's Law implies a smoothing function slightly different from Good-Turing.

Zipf's law is applied in many other areas like ecological systems, genomic data, earthquakes and clinical diagnosis etc. Hill[56] (1974) proposed a modification involving classification of species into a family and then into genera within families. Reed[32, 57, 59, 58] et al. (2002, 2003, 2004) presented models for the size-distribution of forest fires, distribution of family names and size distribution of gene and protein families. It was applied in the distribution of large earthquakes (Sornette[60] et al., 1996). Li[61] et al (2002) applied Zipf's law in importance of genes for cancer classification using micro array data. Tachimori[62] et al. (2002) analyzed the frequency of clinical diagnosis and found that inverse power relationship between the rank order of diagnosis and the frequency of the appearance of these diagnoses exists. They found that both group types have the inverse-power relationship between the rank order of diagnoses and the frequency of the appearance of these diagnoses. (This relationship is called Zipf's law, which is observed in natural language). They found that, "in addition to the clinical diagnoses, medical indices such as average length of hospital stay, frequencies of medical treatments expressed in terms of CD9-CM (International Classification of Disease 9th Revision, Clinical Modification) and medical fees, also follow Zipf's law". He proved that the diagnostic sets based on the doctor's diagnoses followed Zipf's law. They further commented, "The indication that diagnostic sets observe Zipf's law may possibly have major effects on changing the conventional concept of diagnostic frequency rate". There are many more examples like Zipf's law in percolation (Watanabe[63], 1996), in immune system (Burgos[64] et al. 1996), in liquid gas phase transition of nuclei (Ma[65], 1999) and in psychiatric ward (Piqueira[66], 1999).

There have been many applications of the law in natural languages, like English (Miller[67] et al. 1958), Chinese (Rousseau[69] et al, 1992), Voyanich manuscript (Landini[70], 1997), etc. However, there are few applications of the law to random texts. Li[71] (1992) showed that the Zipf's law is applicable to random texts provided it has a very different word structure and length distribution than a natural language. Losee[72] (2001) provided an information theoretic interpretation of Zipf's Law, a power law. Using the regularity noted he suggested that Zipf's Law is a consequence of the statistical dependencies that exist between terms, described here using information theoretic concepts. He found relationships between the frequency-based characteristics of neighboring terms in natural language and the

rank or frequency of the terms. Given the term rank or frequency, he inferred about the entropy, or average information, of a term or a group of terms. The amount of information that one term has about another depends on the rank of one of the terms and of the rank or frequency of the term pair. Using these relationships, he offered a partial explanation of why Zipf's Law occurs as it does. According to Ferrrer-i-Cancho & Sole[73] (2002), "random texts lose the Zipfian shape in the frequency versus rank plot when words are restricted to a certain length, which is not the case in real texts. It is thus clear that monkey languages' partial validity relies on their word length distribution, which we have indicated is unrealistic. These results suggest that future theories of language origin should be able to explain the origin of Zipf's law, instead of using it as a given constraint".

Zipf's law in literatures

Zipf's law postulates that the frequency of occurrence of any word as a function of rank follows a power law with exponent close to unity. It has been applied to many areas like natural languages, monkey-typing texts, web-access statistics, informetrics, finance and business and ecological systems, etc. There is evidence of differences on whether the power law embedded in Zipf's law is actually a Yule distribution (Martindale[1], et al. 1996), lognormal distribution (Perline[2], 1996) or stretched exponential distribution (Laherrere[3], et al, 1998).

Ferrer-i-Cancho & Sole[74,75] (2001) commented that Zipf's law has been a popular achievement of quantitative linguistics. Zipf's appears to be robust. Many models of syntactic communication assume this law. It is an obvious ingredient for any theory of language evolution. A complete theory of language requires a theoretical understanding of its implicit statistical regularities. According to them, "Words in human language interact in sentences in non-random ways, and allow humans to construct an astronomic variety of sentences from a limited number of discrete units. This construction process is extremely fast and robust. The co-occurrences of word in sentences reflect language organization in a subtle manner that can be described in terms of a graph of word interactions".

According to Ferrer-i-Cancho[76] (2005), "Given the apparent universality of Zipf's law and also the enormous differences between all languages on Earth, it is tempting to think that its explanation has nothing to do with language....Zipf's law

for word frequencies could be the manifestation of a complex system operating between order and disorder". According to Miller & Chomsky[77] (1963), "The occurrence of Zipf's law does not constitute evidence of some powerful and universal psychological force that shapes all human communication in a single mould". Zipf's law doesn't manifest at a higher level of semantic cognition where language appears compressed. Therefore, Zipf's law can be rooted in a language structuring process of coding, which adds redundancy necessary for language understanding. Zipf's Law provides a base-line model for expected occurrence of target terms and the answers to certain questions may provide considerable information about its role in the corpus (Steele[78] et al., 1998). Zipf's Law provides a distributional foundation for models of the language learner's exposure to segments, words and constructs, and permits evaluation of learning models (Brent[79], 1997).According to Powers[80] (1998), "Zipf's theory requires effort to be constant independent of frequency, however Information Theory and Psychological experiments both indicate that this ought not to be the case, and that it in fact decreases in a way consistent with an optimal strategy for an unbounded lexicon".

According to Powers[80] (1998), "Zipf considered that the speaker had to build a continuous stream of specified products, that is an ongoing stream of utterances conveying specified meanings, in such a way as to minimize his effort as speaker consistent with effective communication to the hearer, her task being simplified as the relationship between utterances and meanings approached one to one: the work involved in producing a construction consists of the work involved in fetching the tool, which is directly in proportion to the cost of fetching the tool and includes both the mass of the tool, *m*, and the distance, *d*, that it needs to be fetched, given increasing either increases the effort required". According to Ferrer-i-Cancho & Sole[81] (2003), "the early hypothesis of Zipf of a principle of least effort for explaining the law is shown to be sound. Simultaneous minimization in the effort of both hearer and speaker is formalized with a simple optimization process operating on a binary matrix of signal–object associations. Zipf's law is found in the transition between referentially useless systems and indexical reference systems. We strongly suggest that Zipf's law is a hallmark of symbolic reference and not a meaningless feature".

According to Le Quan Ha[82] (2002), "Zipf discovered the law by analyzing manually the frequencies of words in the novel *Ulysses* by James Joyce. It contains a vocabulary of 29,899 different word types associated with 260,430 word tokens". They found that for single words Zipf's law is valid only for high frequency words. Zipf's law performs better on the data containing single word and n-gram phrases combined together. It works for the low frequencies also across languages. According to Li[83] (2002), one of the many phenomenons using Zipf's law pattern is word usage in human languages. The number of times a word is used in written human languages and the frequency of usage are the variables that indulge in a Zipf's type distribution. This phenomenon can also be extended to spoken languages, non-English or non-Latin languages, combination of words, etc.

Smith & Devine[84] (1985) found that legal texts also follows Zipf's law but in a little different manner. They showed that lawyers use more words than other people. Francis & Kucera[85] (1964) applied the Zipf's law to the Brown corpus of 1 million words of American English. A corpus is a body of naturally occurring text, stored in a machine-readable form. Le Quan Ha[82] et al. (2002) analyzed Zipf's law for large corpora in two languages, English and Mandarin. The English corpora used in their experiments are taken from the Wall Street journal and the Mandarin corpus used in their experiments was the TREC Corpus obtained from the People's Daily Newspaper from 01/1991 to12/1993 and from the Xinhua News Agency for 04/1994 to 09/1995 from the Linguistic Data Consortium. Wang[86] (1989) presented Zipf's distribution of Chinese corpus and Wyllys[87] (1981) took a data set of 3907 English words. Sun[88] et al. (1999) proposed a simpler model for estimating the frequency of any same-frequency words and identifying the boundary point between high-frequency words and low-frequency words in a text. The model was based on the maximum ranking method and it ranked words and estimated word frequency with the help of a formulae. They commented, "Studies of word frequency have many interesting and potentially significant applications. For example this model could be used to evaluate a single article or an author's work. Assuming a reasonable level of skill among the writers whose works are the basis for our observations, we can use this model as a benchmark for assessing writer's language skills". According to Pinker & Bloom[89] (1990), "Many authors have pointed out that tradeoffs of utility concerning hearer and speaker needs to appear at many

levels. As for the phonological level, speakers want to minimize articulatory effort and hence encourage brevity and phonological reduction. Hearers want to minimize the effort of understanding and hence desire explicitness and clarity".

Ferrer-i-Cancho & Solé[81] (2003) commented that the effort for the hearer has to do with determining what the word actually means. The higher the ambiguity (i.e. the number of meanings) of a word, the higher the effort for the hearer. Besides, the speaker will tend to choose the most frequent words. The availability of a word is positively correlated with its frequency. Gernsbacher[90] (1994) called this phenomenon as the *word-frequency* effect. Thereafter, the speaker tends to choose the most ambiguous words, which is opposed to the least effort for the hearer. Zipf referred to the lexical tradeoff as the *principle of least effort*. He pointed out that it could explain the pattern of word frequencies, but he did not give a rigorous proof of its validity. Word frequencies obey Zipf's law. If the words of a sample text are ordered by decreasing frequency, the frequency of the k^{th} word, $P(k)$, is given by $P(k) \propto k^{-\alpha}$, with $\alpha \approx 1$(11). According to Balasubramaniyan & Narayan[91] (1996) this pattern is robust and widespread.

According to Deacon[92] (1997), "This might explain why human language is unique with regard to other species but not only so. One-to-one maps between signals and objects are the distinguishing feature of index reference. Symbolic communication is a higher-level reference in which reference results basically from interactions between signals. Zipf's law appears on the edge of the indexical communication phase and implies polysemy. The latter is the necessary (but not sufficient) condition for symbolic reference".

Situngkir[93] reported the statistical observation of Zipf's law to different human languages while the approached corpus is being telling the same things. This is expected to reduce the possible sensitivity to the meaning of the texts and the different stylized statistics are closer to what emerging from the respective structure of language, whether it grammatical or lexical. Interestingly, it has also been showed that Zipfian statistics is robust throughout those raw corpuses analyzed.

According to Stewart[94] (1994), Zipf's law was developed to describe the frequency of word use in documents. He applied Zipf's law to three classic sets of word frequency data: Eldridge's distribution of word usage in four American newspaper

articles, Brugmann's study of four plays in Plautine Latin, and noun frequency in Macaulay's essay on Bacon. He obtained excellent fits to these data sets.

Recent Corpora

Corpus	Size	Domain	Language
NA News Corpus	600 million	Newswire	American English
British National Corpus	100 million	Balanced	British English
EU proceedings	20 million	Legal	10 language pairs
Penn Treebank	2 million	Newswire	American English
Broadcast News		Spoken	7 languages
SwitchBoard	2.4 million	Spoken	American English

Table 2.1: *Some examples of recent corpora*

For more corpora, the Linguistic Data Consortium at http://www.ldc.upenn.edu/ can be visited.

Gelbukh and Sidorov[95] (2001) observed that the coefficients of Zipf law are different for different languages. They illustrated this through English and Russian examples. It is important to reason this as it may have some implications on the nature of language. They further commented that performance of Zipf's law is different in these languages as "Russian is a highly inflective language while English is analytical. Spanish, having "inflectivity" intermediate between Russian and English, showed intermediate results as to the coefficients. The other aspect is that lexical richness of Russian is greater than that of English (and Spanish)". Ferrer-i-Cancho and Sole[74] (2001) showed that the co-occurrence of words in sentences relies on the network structure of the lexicon. They analyzed the properties in depth and commented that human language can be described in terms of a graph of word interactions.

Turner[96] (1997) investigated relationship between vocabulary, text length and Zipf's law. He tried to relate the rate at which previously unused words were added to a text as an author increased its length. The question was whether there exists a relationship between vocabulary and the text length. He has chosen four texts viz. two Shakespearean plays- *Anthony and Cleopatra & Richard III* and two novels

Withering Heights by Emily Bronte, *Sense and Sensibility* by Jane Austin. He found that the rate at which new words are added to the text as its length increases follows a power law. The rate is lower than that from Zipf's law for both novels and plays. The rate came out to be approximately two thirds for plays and a half for novels. He also commented that a deviation of actual text from Zipf's law distribution needs classification and explanation.

Landini[70] (1997) applied Zipf's law in Voynich manuscript (The mysterious manuscript which is still unread). It is still not known that in which language it is written, about its alphabets and abbreviations etc. However it was observed that rank frequency and length frequency are still present in this manuscript. Li[97] (1998) showed that Zipf's law is applicable in random texts also, however such random texts should have a very different words and text length distribution than a natural language. Perhaps, this is the reason for this law appearing in natural languages different from those in random collection of characters. Zipf searched for a principle of least effort that would explain the equilibrium between uniformity and diversity in usage of words. Most others searched for a probabilistic explanation. The burning question still remains- Do we have any new evidence that Zipf's explanation of principle of least effort is more correct than a statistical explanation?

There have been many applications of the law in natural languages, like English (Miller[67] et. al. 1958), Chinese (Rousseau[69] et al. 1992), Voyanich manuscript (Landini[70], 1997), etc. However, there are few applications of the law to random texts. Li[97] (1998) showed that the Zipf's law is applicable to random texts provided it has a very different word structure and length distribution than a natural language.

To investigate more into this area, Saxena[100] et al (2004) selected a random text and tried to find clues on the distribution of rank and frequency. An attempt has been made to evolve a new ranking method, based on tied-ranks and a comparison has been made with the random rank method, deployed by Zipf[99] (1949) and maximum rank method, deployed by Chen & Leimkuhler[98] (1987). According to Mandelbrot[101] (1953), "The monkey language is, in the terminology of fractal geometry, self-similar and grows on infinite trees (any branch of the tree will be identical to the tree itself), thus needing an infinite dictionary. A natural language like English, on the other hand, is a massively geared down system that economizes on entropy in a number of ways, e.g., the interdependence—or redundancy—of

words that seems necessary in order to make a text "meaningful." Most letter combinations (an uncountable set) in English are non-words". However, the random text taken for analysis in this communication is called "random" only because though it is in English, it follows a very subject specific usage of words, e.g. use of hyphenated words. Hence, in this communication, the random text used, differs from monkey typing text by only one virtue, i.e. every word in this random text has a definite meaning.

Chao & D'haeseleer[40] (2001) attempted to find the distribution of Variable length phatic interjectives on the World Wide Web. They found that the number of pages found containing these words would fall off as a power law. However the exponents for length frequency distributions of different interjectives were much larger than –1 predicted by Zipf's law. Parunak[102] (1979) developed a data-analytic technique and applied it on count of words from large texts in Greek, French and English. According to them, "Counted data, whether the number of words in a text or the number of animals of various species in a population, often lacks the usual forms of structure....It is however possible to rank the counts from most frequent to least frequent. The resulting frequency distributions are usually very long tailed and they follow a fairly regular pattern, which is approximated by the rank-size rule". It was found out that word frequency distributions are dependent on word-length.

Sen[103] et al. (1998) investigated the application of Zipf's law on technical writing. They commented that "technical writing differs from literary or ordinary writing in a number of ways. In technical writing, more often than not, each term represents a particular concept which is used again and again whenever the author refers to that concept thus leading to the increase in the frequency of its use". It was found that the LIS writings also follow the Zipf's law when only the textual part of the writing is considered omitting alpha-numeric and alpha-symbolic expressions, abbreviations, heading of illustrations, intra-text references, words figuring within table and keywords.

There are other instances of Zipf's law in natural languages like Dahl[104] (1979) analyzed word frequencies of spoken American (Verbatim). He found that the top twenty words were: I, and, the, to, that, you, it, of, a, know, was, uh, in, but, is, this, me, about, just, don't etc. A similar result has been obtained by Ferrer-i-Cancho & Sole[108] (2001). They commented that the so-called particles, a subset of the function

words (e.g. articles, prepositions and conjunctions) which are used for speeding-up the navigation forms the most frequently occurring words. The top ten were found as "and", "the", "of", "in", "a", "to","s", "with", "by", and "is".

Ridley & Gonzales[105] (1994) analyzed adult speech and applied Zipf's law to small samples of adult speech. Balasubrahmanyan & Naranan[91] (1998) described models for power law relations in Linguistics and Information Science. Sen[103] et al. (1998) conducted a study that indicated that the technical writings such as LIS writings also follow the Zipf's law when only textual part of the writing is considered omitting alpha-numeric and alpha-symbolic expressions, abbreviations, headings of illustrations, intra-textual references, words and figuring within tables, keywords. Egghe[24] (1999) applied this law for multi-word phrases. Prun[106] (1999) illustrated Zipf's conception of language as an early prototype of synergetic linguistics.

Le Quan Ha[82] et al. (2002) found a confirmation of Zipf's law in the extended form. They found that n-gram word phrases as well as single words follow Zipf's law accurately. They verified this result valid for five languages viz. English, Mandarin, Irish, Latin and Vietnamese. Martynyuk[107] (2006) applied Zipf's law on Hindi and Urdu texts. According to Martynyuk[107] (2006), "Statistical regularities are the basis of the structure of the vocabulary of any language or text. Zipf's law is a reflection of a specific property of the organization of human memory, which usually operates with more frequent language units in all cases of the spontaneous use of speech".

References

1. Martindale, Colin., Konopka, Andrzej. K. (1996). Oligonucleotide frequencies in DNA follow a Yule distribution, *Computer & Chemistry*, 20(1):35-38.
2. Perline, Richard. (1996). Zipf's law, the central limit theorem, and the random division of the unit interval, *Physical Review E*, 54(1):220-223.
3. Laherrere, Jean., Sornette, D. (1998). Stretched exponential distributions in Nature and Economy: 'Fat tails' with characteristic scales, *European Physical Journals*, B2:525-539.
4. Reed, W.J. (2001). The Pareto, Zipf and other power laws, *Economics Letters*, 2001, vol. 74, issue 1, pages 15-19.
5. Reed, W.J. (2002). On the rank-size distribution for human settlements, *J Regional Science*, 41:1-17.
6. Reed, W. J., and Hughes, B. D. (2002). From gene families and genera to incomes and internet file sizes: why power-laws are so common in nature. *Phys. Rev. E* 66 067103.
7. Reed, W. J. and Hughes, B. D. (2002). On the size distribution of live genera. *J. Theor. Biol.* 217.
8. Reed, W. J. and Hughes, B. D. (2003). On the distribution of family names. *Physica A* 319:579-590).
9. Bi, Zhiqiang., Faloutsos, C., Korn, F. (2001). The DGX distribution for mining massive, skewed data, *Conference on Knowledge Discovery and Data Mining (KDD)* 2001.
10. Mandelbrot, B.B. (1959). A note on a class of skew distribution function. Analysis and critique of a paper by H.A. Simon, *Information and Control, 2,* 90-99.
11. Simon, H. A. (1955). On a class of skew distribution functions, *Biometrika, 42*:425-440.
12. Simon, H. A. (1960). Some further notes on a class of skew distribution functions, *Information and Control, 3,* 80-88.
13. Mandelbrot, B. B. (1961). Final note on a class of skew distribution functions: analysis and critique of a model due to H.A. Simon, *Information and Control, 4,* 198-216.
14. Hertzel, Dorothy. H. (1987). History of the development of ideas in bibliometrics. *Encyclopedia of Library & Information Sciences, Vol. 42,* Supplement (7), pp 180-219.
15. Wyllys, Ronald. E. (1975). Measuring Scientific Prose with Rank-Frequency ('Zipf') curves: A New Use for an Old Phenomenon, in *Proceedings ASIS 38th Annual Meeting, Inf. Revolution, 12,* 30.

16. Haitun, S. D. (1982). Stationary Scientometric Distributions. Part I. The different approximations. *Scientometrics, 4(1)*, 5-25. Part II. Non-Gaussian nature of scientific activities, *Scientometrics, 4(2)*, 89-104.Part III. The role of Zipf distribution, *Scientometrics, 4(3)*, 181-94.

17. Chai, Kim. (1982). Retrieval Language of Social Sciences and Natural Sciences: A Statistical Investigation, *Journal of the American Society for Information Sciences*, Jan 1982; 33, 1; ABI/Inform Global, Page 3.

18. Thom, James. A., & Zobel, Justin. (1992). A model for word Clustering. *Journal of the American Society for Information Science, 43(9)*, 616-627

19. Witten, I., & Bell, T. (1990). Source models for natural language text. *International Journal of Man-machine Studies, 32,* 545-579

20. Brookes, B.C., (1984). Towards Informetrics: Haitun, Laplace, Zipf, Bradford and the Alvey programme. *Journal of Documentation, Vol. 40*, no. 2, pp120-143.

21. Ivancheva, Ludmila. E. (2001). The Non-Gaussian nature of Bibliometrics and Scientometric distributions: A new approach to interpretation. *Journal of the American Society for Information Science and Technology. 52*, 13, pg. 1100

22. Nicholls, Paul. Travis. (1987). Brief Communication: Estimation of Zipf Parameters. *Journal of the American Society for Information Science (1986-1998);* Nov 1987; 38, 6; pg. 443

23. Apostolos, Georgakis. A. and Li, H. (2003). Document distances using the Zipf distribution and a novel metric. *DML Technical Report,* Department of Applied Physics and Electronics, Umea University, Sweden

24. Egghe, L. (1999). On the law of Zipf-Mandelbrot for multi-word phrases. *Journal of the American Society for Information Science;* Mar 1999; 50, 3; pg. 233

25. Hill, Bruce. M. (1970). Zipf's law and prior distributions for the composition of a population, *Journal of the American Statistical Association, 65*:1220-1232.

26. Makse, Hernan. A., Havlin, Shlomo., Stanley, H. Eugene. (1995). Modeling urban growth patterns, *Nature*, 377:608-612.

27. Krugman, P. (1996). *The Self-Organizing Economy* (Blackwell, Cambridge, MA).

28. Zanette, D. H. and Manrubia, S. C. (1997). Role of intermittency in urban development: a model of large-scale city formation, *Physical Review Letters*, 79:523-526.

29. Manrubia, S. C., Zanette, D. H. (1998). Intermittency model for urban development, *Physical Review E*, 58:295-302.

30. Marsili, Matteo. and Zhang, Yi-Cheng. (1998). Interacting Individuals Leading to Zipf's Law. *Physical Review Letters,* Volume 80, Number 12, 2741-2744.

31. Gabaix, X. (1999). Zipf's law for cities: an explanation, *Quarterly Journal of Economics*, 114:739-767.

32. Reed, W. J. (2002). On the rank-size distribution for human settlements, *J Regional Science*, 41:1-17.

33. Reed, W. J. (2001). The Pareto, Zipf and other power laws, *Economics Letters*, 2001, vol. 74, issue 1, pages 15-19.

34. Knudsen, Thorbjorn. (2001). Zipf's law for cities and beyond: The case of Denmark, *American Journal of Economics and Sociology*, Vol. 60, No. 1.

35. Urzua, Carlos. M. (2000). A simple and efficient test for Zipf's law. *Economics Letters.* 66, 257–260

36. Adamic, Lada. A. & Huberman, Bernardo. A. (2000). The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce*, 1, 5-12.

37. Shi, Lei., Gu, Zhimin., Wei, Lin. and Shi, Yun. (2006). An Applicative Study of Zipf's Law on Web Cache. *International Journal of Information Technology* Vol. 12 No.4 2006

38. Rousseau, Ronald. (2001). Evolution in time of the number of hits in keyword searches on the Internet during one year, with special attention to the use of word Euro, *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, Sydney. Vol. 2, pp 619-627.

39. Lee, Breslau., Pei, Cao., Li, Fan., Graham, Phillips. & Scott, Shenkar. (1999). Web Caching and Zipf-Like Distributions: Evidence and Implications, *IEEE INFOCOM*, Vol. XX. NO- Y.

40. Chao, Dennis., D'haeseleer, Patrik. (2001). The distribution of Variable-Length Phatic Interjectives on the World Wide Web, *UNM Computer Science Department Tech Report* TR-CS-2001-23.

41. Arlitt, Martin. F., & Williamson, Carey. L. (1997). Internet web server: workload characterization and performance implications. *IEEE/ACM Transactions on Networking*, 5(5), 631-645.

42. Glassman, Steve. (1994). A caching relay for the world wide web, In *First International World-Wide Web Conference*, pages 69-76 (May 1994).

43. Crovella, M. E., Bestavros, A. (1997). Self-similarity in World Wide Web traffic: evidence and possible causes, *IEEE/ACM Transactions on Networking*, 5(6):835-846.

44. Carlson, J. M. & Doyle, J. (2000). Highly optimized tolerance: A mechanism for power laws in designed systems, *Physical Review E*, 60(2):1412-1427.

45. Chen, Weishing. & Wu, Tai-His. (1997). A non-homogeneous software reliability model based on Zipf's law. *The International Journal of Quality & Reliability Management.* Vol.14. Issue. 4; pp. 409

46. Champernowne, D. (1953). A model of income distribution, *Economic Journal*, 63:318-351.

47. Mandelbrot, B. B. (1963). New methods in statistical economics, *Journal of Political Economy*, 71:421-440.

48. Mandelbrot, B. B. (1997). *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*, Springer-*Verlag*.

49. Aoyama, H., Souma, W., Nagahara, Y., Okazaki, M. P., Takayasu, H., & Takayasu, M. (2000). Pareto's law for income of individuals and debt of bankrupt companies. *Fractals*. 8(3), 293-300.

50. Reed, W. J. (2001). The Pareto, Zipf and other power laws. *Economics Letters*, 2001, vol. 74, issue 1, pages 15-19.

51. Stanley, M. H. R., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A., Stanley, H. E. (1995). Zipf's plots and the size distribution of firms, *Economics Letters*, 49:453-457.

52. Axtell, Robert. L. (2001). Zipf distribution for US firm sizes. *SCIENCE, 293*.

53. Benee, Valerie. & Oppenheim, Charles. (2004). Does Bradford-Zipf apply to business and management journals in the 2001 Research Assessment Exercise? *Journal of Information Science, 30(5)*, 469-474.

54. Choi, J. S., Kim, Kyungsik., Yoon, S. M., Chang, K. H., & Lee, C. Christopher. (2005). Zipf's Law Distributions in Korean Financial Markets. *Journal of the Korean Physical Society*, Vol. 47, No. 1, July 2005, pp. 171-173

55. Samuelsson, C. (1996). *Relating Turing's Formula and Zipf's Law*. WVLC'96

56. Hill, Bruce. M. (1974). Zipf's law and prior distributions for the composition of a population, *Journal of the American Statistical Association*, 65:1220-1232.

57. Reed, W. J. & McKelvey, K. S. (2002). Power law behaviour and parametric models for the size-distribution of forest fires. *Ecological Modeling*, 150:239-254.

58. Reed, W. J. & Jorgensen, M. (2004). The double Pareto-lognormal distribution - A new parametric model for size distribution. *Com. Stats -Theory & Methods*, Vol. 33, No. 8., 1733-1753.

59. Reed, W. J. and Hughes B. D. (2003). On the distribution of family names. *Physica A*, 319:579-590).

60. Sornette, D., Knopoff, L., Kagan, Y. Y., Vanneste, C. (1996). Rank-ordering statistics of extreme events: application to the distribution of large earthquakes, *Journal of Geophysical Research*, 101(B6):13883-13894

61. Li, W. & Yang, Y. (2002). Zipf's law in importance of genes for cancer classification using micro array data, *Journal of Theoretical Biology*, 219:539-551.

62. Tachimori, Y. & Tahara, T. (2002). Clinical Diagnosis following Zipf's law, *Fractals,* Vol. 10 No. 3, 341-351.

63. Watanabe, M. S. (1996). Zipf's law in percolation, *Physical Review E,* 53(4):4187-4190.

64. Burgos, J. D. & Moreno-Tovar, P. (1996). Zipf-scaling behavior in the immune system, *Biosystems*, 39(3):227-232.

65. Ma, Y. G. (1999). Zipf's law in the liquid gas phase transition of nuclei, *European Physics Journal*, A6:367-371.

66. Piqueira, J. R., Monteiro, L. H., de Magalhaes, T. M., Ramos, R. T., Sassi, R. B. & Cruz, E. G. (1999). Zipf's law organizes a psychiatric ward, *Journal of Theoretical Biology*, 198:439-443.

67. Miller, G. A. & Newman, E. B. (1958). Tests of a statistical explanation of the rank-frequency relation for words in written English. *American Journal of Psychology*, 71, 209-218

68. Rousseau, Ronald. & Zhang, Qiaoqiao. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics. 24(2)*: 201-220.

69. Rousseau, Ronald. & Zhang, Qiaoqiao. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics.* 24(2): 201-220.

70. Landini, G. (1997). Zipf's laws in the Voynich manuscript. http://sun1.bham.ac.uk/G.Landini/evmt/zipf.htm

71. Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory,* 38(6): 1842-1845.

72. Losee, Robert. M.(2001). Term Dependence: A Basis for Luhn and Zipf Models. *Journal of the American Society for Information Science and Technology.* 52 (12), 1019-1025, 2001

73. Ferrer I.Cancho, Ramon. & Sole, Ricard .V. (2002). Zipf's Law and Random Texts. *Advances in Complex Systems*, Vol. 5, No. 1 (2002) 1-6

74. Ferrer I Cancho, Ramon. & Sole, Ricard. V. (2001). The small world of human language, *Proc. R. Soc. Lond.* B (2001) 268, 2261-2265

75. Ferrer I Cancho, Ramon. & Solé, Ricard. V. (2003). Least effort and the origins of scaling in human language, *PNAS* 2003; 100; 788-791; originally published online Jan 22, 2003.

76. Ferrer I Cancho, Ramon. (2005). Zipf's law from a communicative phase transition. In: *European Physical Journal B*, 47; 449–457.

77. Miller, George. A. & Chomsky, Noam. (1963). Finitary models of language users. In: Luce, Robert. D., Bush, Robert. R. & Galanter, Eugene. (Eds.), *Handbook of Mathematical Psychology*, vol. 2. New York: Wiley, 419–491.

78. Steele, R. & Powers, D. M. W. (1998). *Evolution and Evaluation of Document Retrieval Queries.*

79. Brent, M. R. (1997). Toward a Unified Model of Lexical Acquisition and Lexical Access. *Journal of Psycholinguistic Research* 26:363-375.

80. Powers, David. M.W. (1998). Applications and Explanations of Zipf's Law. In Powers, D. M.W. (ed.) NeMLaP3/CoNLL98: *New Methods in Language Processing and Computational Natural Language Learning*, ACL, pp 151-160.

81. Ferrer I Cancho, Ramon. & Solé, Ricard. V. (2003). Least effort and the origins of scaling in human language, *PNAS* 2003; 100; 788-791; originally published online Jan 22, 2003

82. Le Quan, Ha., Sicilia-Garcia E. I., Ming, J. & Smith, F. J. (2002). Extension of Zipf's Law to Words and Phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, pages 315-320, Taipei, Taiwan.

83. Li, Wentian. (2002). Zipf's Law Everywhere, *Glottometrics*, 5, 2002, 14-21

84. Smith, F. J. & Devine, K. (1985). Storing and Retrieving Word Phrases *Information Processing & Management*, Vol. 21, No. 3, pp 215-224.

85. Francis, W. N. & Kucera, H. (1964). *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English*, for use with Digital Computers Department of Linguistics, Brown University, Providence, Rhode Island

86. Wang, C. (1989). Zipf's distribution of Chinese corpus, *Information Sciences*, 10, 1-8

87. Wyllys, Ronald, E. (1981). Empirical and Theoretical Bases of Zipf's Law. *Library Trends* 30(1) (Summer 1981): 53-64

88. Sun, Qinglam., Shaw, D. & Davis, C. H. (1999). A model for estimating the occurrence of same frequency word and the boundary between the high and low frequency words in texts, *Journal of the American Society for Information Science*, Mar 1999: 50, 3

89. Pinker, S. & Bloom, P. (1990). Natural language and natural selection, *Behav. Brain Sci.* 13, 707–784.

90. Gernsbacher, M. A. ed. (1994). *Handbook of Psycholinguistics.* Academic, San Diego.

91. Balasubrahmanyan, V. K. & Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, 3:3, 177–228.

92. Deacon, T. W. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain,* Norton & Company, New York.

93. Situngkir, Hokky. An Observational Framework to the Zipf's Analysis among Different Languages Studies to Indonesian Ethnic Biblical Texts.

94. Stewart, J.A. (1994). The Poisson-Lognornal model for bibliometric/scientometric distributions, *Information Processing & Management,* Vol 30, No.2, pp 239-251.

95. Gelbukh, Alexander. & Sidorov, Grigori. (2001). Zipf and Heaps Laws Coefficients Depend on Language, *Proc. CICLing-2001*, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335.

96. Turner, C. R. (1997). Relationship between vocabulary, text length, and Zipf's law, http://www.btinternet.com/~g.r.turner/ZipfDoc.htm

97. Li, W. (1998). Letter to the editor: Zipf's law in the structure and Evolution of Languages, in *Complexity*, 3(5): 9-10

98. Chen, Ye-Sho. & Leimkuhler, Ferdinand. F. (1987). Analysis of Zipf's Law: An index approach. *Information Processing and Management.* 23(3): 171-182.

99. Zipf, G. K. (1949). *Human Behavior and the principle of least effort,* Cambridge, MA: Addison-Wesley Press

100. Saxena, Anurag., Jauhari, Monika. & Gupta, B. M. (2007). Zipf's Law in a Random Text from English With a New Ranking Method, *DESIDOC Bulletin of Information Technology*, Vol. 27, No. 4, July 2007, pp. 51-58

101. Mandelbrot, B. B. (1953). An informational theory of the statistical structure of languages, in *Communication Theory*, ed. W. Jackson (Butterworth, 1953), pp. 486-502.

102. Parunak, Anita. (1979). Graphical analysis of ranked counts (of words), *Journal of the American Statistical Association,* Volume 74, No- 365.

103. Sen, B. K., Khong, Wye. Keen., Lee, Soo Hoon., Lim Bee. Ling., Abdullah, Mohd. Rafae,, Ting, Chang. Nguan,, Wee, Siu, Hiang. (1998). Zipf's law and writings on LIS. *Malaysian Journal of Library & Information Science*, 3(2). 93-98.

104. Dahl, H. (1979). *Word Frequencies of Spoken American* (Verbatim).

105. Ridley, D. R. & Gonzales, E. A. (1994). Zipf's law extended to small samples of adult speech, *Percept. Mot. Skills*, 79:153-154.

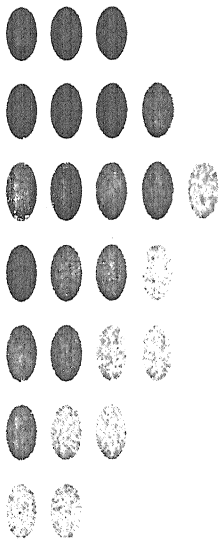
106. Prun, Claudia. (1999). G.K. Zipf's conception of language as an early prototype of synergetic linguistics, *Journal of Quantitative Linguistics*, 6(1)

107. Martynyuk, Stanislav. (2006). Statistical Approach to the Debate on Urdu and Hindi. *The Annual of Urdu Studies.*

108. Ferrer-i-Cancho, Ramon. & Sole, Richard. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited". *Journal of Quantitative Linguistics* 8:165-174.

Chapter 3

Research Methodology



Research Methodology

Objectives

The objectives of the present work are:

- To find the interrelationships between the rank and the frequency of a word in selected literatures.
- To test whether the Zipf's law can be applied in these literatures.
- To do an inter-literature comparison of the applicability of Zipf's law.
- To do mathematical modelling & validation of the model through the collected data.

Hypothesis

The hypotheses of this work are that

1. The principle of least effort is a universal phenomenon,
2. All writers would follow an economy in the use of words irrespective of the language concerned,
3. The rank-frequency distribution of words would be similar in all languages.

The Data

For inter-literature comparison of the applicability of Zipf's law, we have selected the following set of texts from diverse literatures. Thus, we have selected the following 31 sets of text from diverse literatures.

- Computer Science: A text of 10,043 words from a computer science "Operating System - Concepts and Design", by Milenkovic[1], Second edition, 1997 (Tata McGraw Hill, New Delhi).

- **Hindi literature:** The IIT Kanpur's e-text of roman version of "Eidgaah" by Munshi Prem Chand[2]. (http://www.munsipremchand.iitk.ac.in/auther.html). This website has been built as part of a larger effort to create a series of websites based on Indian philosophical texts. This website has been built under a project in the Department of Computer Science & Engineering at the Indian Institute of Technology Kanpur. (File: eidgaah.txt)

- **English**: The Project Gutenberg e-text of "Aladdin and the Wonder Lamp", a "public domain" work distributed by Professor Michael S. Hart[3] through the Project Gutenberg Association. Project Gutenberg is the oldest producer of free e-books on the Internet (http://www.gutenberg.org/). (File: aladdin eng.txt)

- **German:** The Project Gutenberg e-text of "Aladdin und die Wunderlampe", by Ludwig Fulda[4], with original illustration by Max Liebert. Project Gutenberg is the oldest producer of free e-books on the Internet (http://www.gutenberg.org/).(File: aladdin ger.txt)

- **Library Science:** The Project Gutenberg e-text of "The Library", by Andrew Lang[5] #20 in our series by Andrew Lang, December, 1999 (File: librarys.txt)

- **Sanskrit:** The Project Gutenberg e-Book of "Sri Vishnu Sahasranaamam", by Unknown. It is in Sanskrit and character set encoding is US-ASCII. This E-text was transcribed by N. Srinivasan & Karthik Krishnan[6] and formatted by Maitri Venkat-Ramani. This e-text can be transliterated in Sanskrit using the ITRANS processing tool at the following location. (File: sanskritwork.txt).
  http://sanskrit.gde.to/processing_tools/processing_tools.html

- **A Language of India:** For this portion we have taken an e-text from the English version of the collection of Ghazal "Bisat-e-Hyder" by Hyder Zaheer Ansari Hyder[7]. (http://www.bisatehyder.indiaaccess.com/) (File: urdu.txt)

- **A book on Thesaurus & their meanings or a dictionary:** The Project Gutenberg E-text of "Mr. Honey's Small Business Dictionary (English-German)" by Winfred Honig[8]. Mr. Honey (Winfred Honig) compiled English/German dictionaries for almost 3 decades to provide his colleagues and students with samples of the language

of business and highlight the need for special dictionaries covering the special language used in different branches of the industry. These wordlists are now fed into the LEO Online Dictionary (http://dict.leo.org) and the DicData Online Dictionary (http://www.dicdata.de) (File: Eng-ger-busDictionary.txt).

Apart from these we have taken many texts over a period of time and some poplular e-texts from Project Guttenberg database and Infomotions etc.

- **E-texts over time:** Public domain electronic texts (e-texts) [9] in the areas of American and English literature as well as Western philosophy are taken in this category. These were "classic" texts that have stood the test of time. They also encompass a huge time period- as far back as 400BC to the present. (http://www.infomotions.com/etexts/)
- **Popular e-texts:** Popular e-texts like "365 Foreign Dishes", "The Arabian Nights Entertainments", "The Arctic Queen" and "The Atomic Bombings of Hiroshima and Nagasaki" were also taken to investigate the relationship.

The following is a description of these texts:

no	Text from	Title	File Name	No- of Words	Appendix
1	Computer Science	Operating System - Concepts and Design by Milan Milenkovic	comsc.txt	10043	14
2	Hindi literature	Eidgaah By Munshi Prem Chand	eidgaah.txt	4951	18
3	English	Aladdin and the Wonder Lamp	aladdin eng.txt	5319	2
4	German	Aladdin und die Wunderlampe	aladdin ger.txt	17686	3
5	Library Science	The Library by Andrew Lang	librarys.txt	37498	23
6	Sanskrit	Sri Vishnu Sahasranaamam	sanskritwork.txt	1411	27
7	Urdu	Bisat-e-Hyder by Hyder Zaheer Ansari Hyder.	urdu.txt	4035	32
8	Dictionary/ Thesaurus	Mr. Honey's Small Business Dictionary (English-German) by Winfred Honig	Eng-ger-busDictionary.txt	21843	16,17

Table 3.1: *Description of first set of documents*

The descriptions of other texts are as follows:

no	Text from	Title	No- of Words	Appendix
1.	American Literature 1700-1799	The Autobiography of Benjamin Franklin	68157	19
2.	American Literature 1800-1899	Autobiography by Thomas Jefferson 1743 – 1790 (With the Declaration of Independence)	40648	21
3.	American Literature 1900-1999	Tom Sawyer, Detective By Mark Twain from "The Writings of Mark Twain, Volume XX	24486	30
4.	English Literature 700-799	Beowulf, from The Harvard Classics, Volume 49	27129	11
5.	English Literature 1200-1299	The Canterbury Tales by Geoffrey Chaucer	99403	13
6.	English Literature 1500-1599	Romeo and Juliet by Shakespeare	26784	29
7.	English Litera ure 1600-1699	The Pilgrim's Progress, by John Bunyan	57122	9
8.	English Literature 1600-1699	Hamlet by Shakespeare	33098	28
9.	English Literature 1700-1799	The Wrongs of Woman by Mary Wollstonecraft	45874	31
10	English Literature 1800-1899	A Christmas Carol by Charles Dickens	21818	15
11	English Literature 1800-1899	Endymion: A Poetic Romance by John Keats	31962	22
12	English Literature 1900-1999	Peter Pan by James M. Barrie	47885	10
13	Western Philosophy 400BC-301BC	Meteorology by Aristotle	43470	6
14	Western Philosophy 100BC-1BC	On The Nature of Things by Titus Lucretius Carus	75386	25
15	Western Philosophy 400-499	Confessions and Enchiridion by Saint Augustine	176014	8
16	Western Philosophy 1600-1699	Concerning Civil Government, Second Essay- An essay concerning the true original extent and end of Civil Government, by John Locke, Chapter I	53786	24
17	Western Philosophy 1700-1799	A Treatise Concerning The Principles of Human Knowledge by George Berkeley	36342	12
18	Western Philosophy 1800-1899	The Subjection of Women by John Stuart Mill	45240	26
19	Western Philosophy 1900-Present	A Young Girl's Diary Prefaced with a Letter by Sigmund Freud	72133	20
20	Popular	365 Foreign Dishes, by Unknown	27891	1
21	Popular	The Arabian Nights Entertainments, by Anonymous	90768	4
22	Popular	The Arctic Queen, by Unknown	16703	5
23	Popular	The Atomic Bombings of Hiroshima and Nagasaki by The Manhattan Engineer District	25341	7

Table 3.2: *Description of second set of documents*

The Software

We have tried much software for calculating the word frequency from the text. We searched the World Wide Web (www) for freeware or shareware, which can do this work. We found four major software. These were Hermetic Word Frequency Counter 5.32, Textanz Word and Phrase Frequency Counter v.1.3, Fore Words Pro 1.2.0.41 and TextSTAT. We tried to analyze various text files with these software. The first three software calculated the frequencies but since we were using the demo version, we faced a major limitation of not been able to transfer the output to a file. We therefore switched to TextSTAT[10] which is completely free software. Thus the Software for calculating the word frequency from the texts used in this work is "TextSTAT". Shown below is a screen shot of TextSTAT.

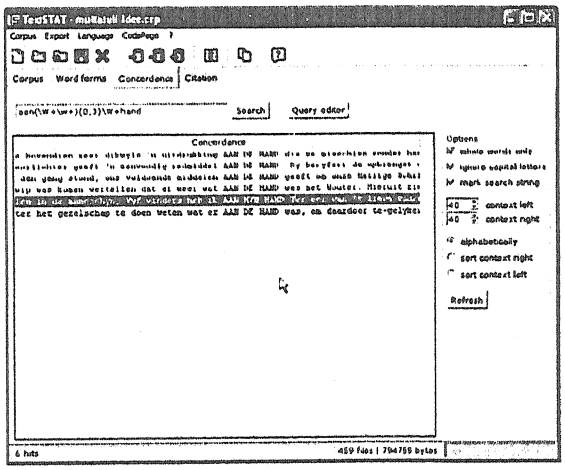


Figure 3.1: *A screen shot of TextSTAT*

TextSTAT is a simple program for the analysis of texts made by Free University of Berlin. It produces word frequency lists and concordances from ASCII/ANSI texts, MS Word and HTML files. TextSTAT can be downloaded from the website http://www.niederlandistik.fu-berlin.de/textstat/software-en.html.

All unique words were ranked at random according to their frequency of occurrence in a decreasing order. Different ranks were assigned to each of them according to Zipf's approach of random-ranks.

Ranking Method

Zipf[12](1949) used random rank approach i.e. words arranged in decreasing order of frequency and ranks allotted in ascending order. In this way the word with maximum frequency will get rank 1 and so on. This leads to steps for large values of rank. This is one of the disadvantages with the random rank method. Chen and Leimukuhler[11] (1987) had overcome this problem, by using the maximum rank for all the words with the same rank. Also, their method helped in preserving the convertibility between frequency-rank distribution & frequency-count distribution and vice-versa, which was not possible in random rank approach. Another method proposed by us is based on the concept of "ties", which means, that if two observations are tied, i.e. they have the same frequency then they should be assigned the ranks according to the average of their random ranks. This was done in order to stabilize the product of frequency and rank. This method is demonstrated with the data for the computer science literature. However, in all other texts the random rank approach of Zipf will be applied.

Analysis

All unique words were ranked at random according to their frequency of occurrence in a decreasing order. Different ranks were assigned to each of them according to Zipf's approach of random-ranks. We then found out the rank frequency g(r) i.e. the number of words of the same rank. This was done in order to obtain the product r x g(r). Here, r is the word rank and g (r) is the rank frequency i.e. the number of words of the same rank.

Microsoft Excel has been used extensively to "sort" the data in the first place and "advanced filter" feature of the Excel is used to filter out the unique frequencies.

A brief description of these features is given below:

Sort: Sort rows in ascending order based on the contents of one column. Alphabets are sorted in ascending alphabetic order and numbers are sorted from lowest to highest value.

Advanced Filter: Advanced filter criteria can include multiple conditions applied in a single column, multiple criteria applied to multiple columns, and conditions created as the result of a formula. It can filter out unique records from a column.

Count If: Counts the number of cells within a range that meet the given criteria.

Once the Zipf an data has been obtained for the various files we have calculated the log (Base 10) values for the rank and rank frequency. Regression analysis and curve-fitting was done on this data. A linear fit was done in order to find the applicability of Zipf-Mandelbrot law. Mandelbrot assumed that the aim of language is to transmit the most information per symbol with the least effort. He proposed the following relationship:

$$f = k(r + c)^{-\theta}$$

Where, f is the frequency and r is the rank of the word; c and θ are constants. Here, c improves the fit for small r and the exponent θ improves the fit for large r. A data follows Zipfian distribution if the exponent θ remains close to -1.

We have used various statistical packages like SPSS, Minitab and Curve Expert to carry out these analyses on the selected texts.

A note on Nonlinear Regression

Nonlinear regression fits a mathematical model to data. A mathematical model is a simple description of a state or process. a model can helps in designing better experiments and make sense of the results. According to Levins[13] (1966), "A mathematical model is neither a hypothesis nor a theory. Unlike scientific hypotheses, a model is not verifiable directly by an experiment. For all models are both true and false.... The validation of a model is not that it is "true" but that it generates good testable hypotheses relevant to important problems". When one fits a model to data, one obtains best-fit values that can be interpreted in the context of the model.

Some programs automatically fit data to hundreds or thousands of equations and then present with the equation(s) that fit the data best. The goal of nonlinear regression is to adjust the values of the variables in the model to find the curve that best predicts Y from X. More simply, the goal is to find the curve that comes closest to the points. To ascertain this, the regression procedure minimizes the sum of the squares of the vertical distances

of the points from the curve. For this reason, linear and nonlinear regressions are sometimes called least squares methods. Some nonlinear regression problems can be linearized by a suitable transformation of the model formulation. For example, consider the nonlinear regression problem (ignoring the error):

Say for the exponential family let us take a model of the form $y = a\exp^{bx}$. If we take a logarithm of both sides, it becomes $\log y = \log a + bx$. Now only estimation of the unknown parameters by a linear regression of log(y) on x is required.

In Curve Expert[14], the nonlinear models have been divided into families based on their characteristic behavior. These families and their members are enumerated below:

Exponential Family

Exponential models have the exponential or logarithmic functions involved. They are generally convex or concave curves, but some models in this group are able to have an inflection point and a maximum or minimum.

Exponential:	y=a*exp(b*x)
Modified Exponential:	y = a*exp(b/x)
Logarithm:	y = a+b*ln(x)
Reciprocal Logarithm:	y = 1/(a+b*ln(x))
Vapor Pressure Model:	y = exp(a+b/x+c*ln(x))

Power Family

The Power Family involves raising one or more parameters to the power of the independent variable, or raising the dependent variable to the power of a given parameter. This family is generally a set of convex or concave curves with no inflection points or maxima/minima.

Power Fit:	$y= a*x^b$
Modified Power:	$y = a*b^x$
Shifted Power:	$y = a*(x-b)^c$

Geometric: $y = a*x^{(b*x)}$

Modified Geometric: $y = a*x^{(b/x)}$

Root Fit: $y = a^{(1/x)}$

Hoerl Model: $y = a*(b^{x})*(x^{c})$

Yield-Density Models

The yield-density models are widely used, especially in agricultural applications. These models historically have been used to model the relationship between the yield of a crop and the spacing or density or planting. Essentially two types of response are observed in practice: the "asymptotic" and "parabolic" yield-density relations. If the response is such that as density (x) increases, but the yield (y) approaches a fixed value, the relationship is asymptotic. If the response is such that there is a distinct optimum as the density increases, the relationship is parabolic. Of course, these types of relationships occur commonly in other scientific areas; therefore, this family of models is very useful.

Reciprocal Model: $y = 1 / (a + bx)$

Reciprocal Quadratic: $y = 1 / (a + bx + cx^2)$

Bleasdale Model: $y = (a + bx)^{(-1/c)}$

Harris Model: $y = 1 / (a + bx^c)$

Growth Family

Growth models are characterized by a monotonic growth from some fixed value to an asymptote. These models are most common the engineering sciences.

Exponential Assoc (2): $y = a*(1-\exp(-bx))$

Exponential Assoc (3): $y = a*(b-\exp(-cx))$

Saturation Growth: $y = ax / (b + x)$

Sigmoidal Family

Processes producing sigmoidal or "S-shaped" growth curves are common in a wide variety of applications such as biology, engineering, agriculture, and economics. These

curves start at a fixed point and increase their growth rate monotonically to reach an inflection point. After this, the growth rate approaches a final value asymptotically. This family is actually a subset of the Growth Family, but is separated because of their distinctive behavior.

Gompertz Model:	$y = a * \exp(-\exp(b - cx))$
Logistic Model:	$y = a / (1 + \exp(b - cx))$
Richards Model:	$y = a / (1 + \exp(b - cx))^{(1/d)}$
MMF Model:	$y = (ab + cx^d)/(b + x^d)$
Weibull Model:	$y = a - b*\exp(-cx^d)$

Miscellaneous Family

As with many things in life, some things just don't fit into nice categories. The miscellaneous family is the one in which these "different" nonlinear regression models live.

Sinusoidal Fit:	$y = a + b*\cos(c*x + d)$
Gaussian Model:	$y = a*\exp((-(x - b)^2)/(2*c^2))$
Hyperbolic Fit:	$y = a + b/x$
Heat-Capacity Model:	$y = a + bx + c/x^2$
Rational Function:	$y = (a + bx) / (1 + cx + dx^2)$

According to Hyams, "Given a set of data points, often called "observations," a common need is to condense the data by fitting it to a model in the form of a parametric equation. This "model equation" can be anything that the user desires -- it can range from a simple polynomial to an extremely complex model with many parameters". One should try to uncover the underlying law that data offers and then select appropriate model. Regression is one of the several techniques of data modeling. Regression ensures that the "merit function", which measures the disagreement between the data and the model, is minimized with respect to the adjustment of the model parameters.

If one take a linear model of the form $y = a_1 X_1(x) + a_2 X_2(x) + ... + a_n X_n(x)$, where $X_i(x)$ could be non-linear function also but a_i are linear. Linear regression can be used to minimize the difference between the model and data. The merit function in this case would be

$$S = \sum_{i=1}^{n} \left(y_i - \sum_1^n a_j X_j(x) \right)^2$$

This has to be minimized the parameters a_k are obtained in this way.

In the case of non-linear regressions, the following method is used as per the documentation of the program Curve Expert. The program uses the Levenberg-Marquardt method to solve nonlinear regressions. This method combines the steepest-descent method and a Taylor series based method to obtain a fast, reliable technique for nonlinear optimization. Neither of the above optimization methods are ideal all of the time; the steepest descent method works best far away from the minimum and the Taylor series method works best close to the minimum. The Levenberg-Marquardt (LM) algorithm allows for a smooth transition between these two methods as the iteration proceeds.

In general, the data modeling equation (with one independent variable) can be written as follows:

$$y = y(x; \bar{a})$$

The above expression simply states that the dependent variable '$\underline{y}$' can be expressed as a function of the independent variable '$\underline{x}$' and vector of parameters '$\underline{a}$' of arbitrary length. Note that using the ML method, any nonlinear equation with an arbitrary number of parameters can be used as the data modeling equation. Then, the "merit function" we are trying to minimize is

$$\chi^2(\bar{a}) = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i; \bar{a})}{\sigma_i} \right)^2$$

Where N is the number of data points, x_i denotes the x data points, y_i denotes the y data points, s_i is the standard deviation (uncertainty) at point i, and $y(x_i,a)$ is an arbitrary nonlinear model evaluated at the i^{th} data point. This merit function simply measures the agreement between the data points and the parametric model; a smaller value for the merit function denotes better agreement. Commonly, this merit function is called the chi-square.

A note on Project Gutenberg e-text & e-books

Project Gutenberg[16] is the first and largest single collection of free electronic books, or e-books. Michae Hart, founder of Project Gutenberg, invented e-books in 1971 and continues to inspire the creation of e-books and related technologies today. Project Gutenberg began in 1971 when Michael Hart was given an operator's account with $100,000,000 of computer time in it by the operators of the Xerox Sigma V mainframe at the Materials Research Lab at the University of Illinois.

Hart announced that the greatest value created by computers would not be computing, but would be the storage, retrieval, and searching of what was stored in our libraries. Project's eventual goal is to provide Public Domain e-text editions a short time after they enter the Public Domain. Of course, the period before a copyrighted work entered the Public Domain was extended from 28 years (with a 28 year extension available) to 50 years more than the life of the author

The Project Gutenberg Philosophy is to make information, books and other materials available to the general public in forms a vast majority of the computers, programs and people can easily read, use, quote, and search. There are three portions of the Project Gutenberg Library, basically be described as:

- Light Literature; such as Alice in Wonderland, Through the Looking-Glass, Peter Pan, Aesop's Fables, etc.
- Heavy Literature; such as the Bible or other religious documents, Shakespeare, Moby Dick, Paradise Lost, etc.
- References; such as Roget's Thesaurus, almanacs, and a set of encyclopedia, dictionaries, etc.

A note on the IIT Kanpur's e-text

In 1990-92 Professor R.M.K. Sinha[15] conceptualized design of a Machine Aided Translation system for translation from English to Indian Languages. This system was named as ANGLABHARTI and the underlying methodology named as ANGLABHARTI Technology or ANGLABHARTI Approach.

ANGLABHARTI represents a machine-aided translation methodology specifically designed for translating English to Indian languages. Indian languages are relatively of free word-order. Instead of designing translators for English to each Indian language, Anglabharti uses a pseudo-interlingua approach. It analyses English only once and creates an intermediate structure called PLIL (Pseudo Lingua for Indian Languages). This is the basic translation process translating the English source language to PLIL with most of the disambiguation having been performed. The PLIL structure is then converted to each Indian language through a process of text-generation.

During 1995-97, Department of Electronics, Govt. of India, sanctioned a grant-in-aid for implementation of the project titled "Machine Aided Translation from English to Hindi for standard documents (domain of Public Health Campaign) based on ANGLABHARTI approach". In 1995-96, IITK also designed and developed an Example-based approach for Machine Aided Translation for similar (Indian languages) and dissimilar (English and Indian Languages) under the leadership of Professor R.M.K. Sinha. This approach has been named as ANUBHARTI approach.

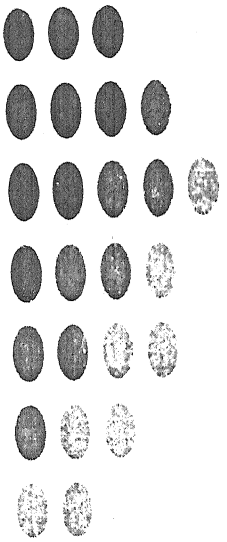
Currently, AnglaHindi, the English to Hindi MAT based on Anglabharti methodology, which accepts unconstrained text, has already been made available to the users and is very well received. AnglaUrdu which is based on AnglaHindi has also been demonstrated. HindiAngla, the Hindi to English MAT based on Anubharti methodology, has been demonstrated for simple sentences and further work is going on to handle compound and complex sentences.

References

1. Milenkovic, Milan. (1997). *Operating System - Concepts and Design*, Second edition. Tata McGraw Hill, New Delhi.
2. The IIT Kanpur's e-text of roman version of "Eidgaah" by Munshi Prem Chand (http://www.munsipremchand.iitk.ac.in/auther.html)
3. The Project Gutenberg e-text of "Aladdin and the Wonder Lamp" (http://www.gutenberg.org/)
4. The Project Gutenberg e-text of "Aladdin und die Wunderlampe", by Ludwig Fulda (http://www.gutenberg.org/)
5. The Project Gutenberg e-text of "The Library", by Andrew Lang #20 (http://www.gutenberg.org/)
6. The Project Gutenberg e-Book of "Sri Vishnu Sahasranaamam" (http://www.gutenberg.org/)
7. The e-text from the collection of Ghazal "Bisat-e-Hyder" by Hyder Zaheer Ansari Hyder. (http://www.bisatehyder.indiaaccess.com/)
8. The Project Gutenberg E-text of "Mr. Honey's Small Business Dictionary (English-German)" by Winfred Honig.
9. The Public domain electronic texts (e-texts) in the areas of American and English literature as well as Western philosophy (http://www.infomotions.com/etexts/)
10. For downloading TextSTAT, the website http://www.niederlandistik.fu-berlin.de/textstat/software-en.html was used.
11. Chen, Ye-Sho. & Leimkuhler, Ferdinand. F. (1987). Analysis of Zipf's Law: An index approach. *Information Processing and Management.* 23(3): 171-182.
12. Zipf, G. K. (1949). *Human Behavior and the principle of least effort,* Cambridge, MA: Addison-Wesley Press
13. Levins, R. (1966). *Am. Scientist.* 54:421-31
14. Hyams, Daniel. For information about CurveExpert 1.3, a comprehensive curve fitting system for Windows and for Curve Expert help documentation (http://curveexpert.webhop.net)
15. Sinha, R. M. K. (2007). For information about IIT Kanpur's e-text. http://www.cse.iitk.ac.in/users/langtech/anglabharti.htm
16. http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart for information about Project Guttenberg e-text & e-books.

Chapter 4

Analysis



Analysis

We have analysed the different documents collected individually. We present the findings obtained by the analysis in the following sections.

Section 1: Zipf's Law in Computer Science Literature

First Analysis: *Whether Zipf's law is applicable on random text in English language from Computer Science literature.*

Zipf's law postulates that the frequency of occurrence of any word as a function of rank follows a power law with exponent close to unity. It has been applied to many areas like natural languages, monkey-typing texts, web-access statistics, informetrics, finance and business and ecological systems, etc. There is evidence of differences on whether the power law embedded in Zipf's law is actually a Yule distribution (Martindale[1], et al. 1996), lognormal distribution (Perline[2], 1996) or stretched exponential distribution (Laherrere[3], et al, 1998). There have been many applications of the law in natural languages, like English (Miller[4] et. al. 1958), Chinese (Rousseau[5] et al, 1992), Voyanich manuscript (Landini[6], 1997), etc. However, there are few applications of the law to random texts. Li[7] (1998) showed that the Zipf's law is applicable to random texts provided it has a very different word structure and length distribution than a natural language.

To investigate more into this area, we have selected a random text from Computer Science literature and have tried to find clues on the distribution of rank and frequency. An attempt has been made to evolve a new ranking method, based on tied-ranks and a comparison has been made with the random rank method, deployed by Zipf[8] (1949) and maximum rank method, deployed by Chen & Leimkuhler[9] (1987). According to Mandelbrot[10] (1953), "The monkey language is, in the terminology of fractal geometry, self-similar and grows on infinite trees (any branch of the tree will be identical to the tree itself), thus needing an infinite dictionary. A natural language like English, on the other hand, is a massively geared down system that economizes on entropy in a number of

ways, e.g., the interdependence—or redundancy—of words that seems necessary in order to make a text "meaningful." Most letter combinations (an uncountable set) in English are non-words". However, the random text taken for analysis in this communication is called "random" only because though it is in English, it follows a very subject specific usage of words, e.g. use of hyphenated words. Hence, in this communication, the random text used, differs from monkey typing text by only one virtue, i.e. every word in this random text has a definite meaning.

Methodology

To study the application of Zipf's law and the performance of the new ranking method on random texts, the authors have taken a text from a computer science " Operating System - Concepts and Design", by Milan Milenkovic[11] , Second edition, 1997 (Tata McGraw Hill, New Delhi).

Word Example	Length	Frequency
A	1	205
AN	2	1765
CAD	3	1580
AREA	4	1100
LOGIN	5	730
DESIGN	6	856
ADDRESS	7	1076
LANGUAGE	8	844
INTERVALS	9	775
CONCURRENT	10	423
UTILIZATION	11	285
ABSTRACTIONS	12	165
COMMUNICATION	13	84
USER-SPECIFIED	14	37
CHANGE_PASSWORD	15	40
REMOTE-PROCEDURE	16	54
MEMORY-MANAGEMENT	17	7
PROGRAMMER-DEFINED	18	5
ADDRESS-TRANSLATION	19	4
LOWER-PRIORITY-BASED	20	3
COMPUTATION-INTENSIVE	21	2
TRANSACTION-PROCESSING	22	1
APPLICATION-PROGRAMMING	23	1

Table 4.1: *Description of words according to length & frequency in Computer Sc Literature*

The authors have counted the frequency of occurrence of each unique word in the text, and found 1775 unique or different words out of a total of 10,043 words in the full text. It was observed that the words of less than 9 characters in length were extensively used. However, one striking characteristic of computer science literature was the use of hyphenated words, which makes the word length vary over a large range. One can easily see from the table below that after words having 13 characters, there are a series of hyphenated words.

Use of hyphenated words can be taken as a special characteristic of the text taken, i.e. the computer science literature. It would thus be interesting to investigate the rank and frequency relationship as propounded by Zipf and other scientists in such a text. The authors have intentionally kept the hyphenated words as they are. One can also see that hyphenated words are typical in describing the very specific nature of the meaning they convey in the concerned literature. Some of them are the commands given to the computer to perform specific tasks.

All unique words were arbitrarily ranked according to their frequency of occurrence in a decreasing order. Words, which shared the same frequency, were arranged alphabetically and different ranks were assigned to each of them according to Zipf's approach of random-ranks. Thus, the words "able" got the rank(r) 868 and the word "writes" got the rank(r) 1775. One can see that two words contributing 1 occurrence each are assigned random ranks 868 and 1775, respectively according to Zipf's random rank approach. This leads to steps for large values of rank. This is one of the disadvantages with the random rank method. Chen and Leimukuhler[9] (1987) had overcome this problem, by using the maximum rank for all the words with the same rank. Also their method helped in preserving the convertibility between frequency-rank distribution & frequency-count distribution and vice-versa, which was not possible in random rank approach.

Another method proposed by us is based on the concept of "ties", which means, that if two observations are tied, i.e. they have the same frequency then they should be assigned the ranks according to the average of their random ranks. This was done in order to stabilize the product r x g(r), especially in the last rank-range. Here, r is the word rank and g (r) is the rank frequency i.e. the number of words of the same rank.

Analysis and Results

The authors had expected that the new ranking procedure based on "ties" would be able to minimize the dispersion of the product r x g(r) in all the rank range due to a simple logic that the maximum rank would always be greater than the average rank. A preliminary analysis of the product r x g(r) is as follows:

Rank range	r x g(r) by Maximal Rank Method			r x g(r) by Tied Rank Method		
	Max	Min	Std. Dev	Max	Min	Std. Dev
1-10	1240	553	227.45	1377	553	227.4
11-51	1485	1239	57.79	1501	1239	62.85
52-99	1548	1352	56.23	1503	1352	46.15
108-228	1596	1512	30.99	1503	1456	16.29
276-1775	1775	1656	40.47	1538	1321.5	83.79

Table 4.2: *Rank frequency relationships in different rank methods*

It can be seen from the above table that the r x g(r) is distributed with fairly less variability but for the rank-range (1-10). This is due to the fact that observation with rank 1 is a clear outlier. If we delete that observation from our calculation of standard deviation then the variability substantially reduces and comes down to 104.61 instead of 227.45. Also an interesting observation is that method of tied rank shows the same variability in the rank range (1-51), performs better in the rank range (52-228) and performs badly in the rank range (276-1775) when compared to the maximal rank method.

Statistical Measure	Ranking Procedure		
	Zipf	*Chen*	*Tied*
Std. Dev	223.76	99.14	86.47
Mean	1393.93	1718.16	1393.93
% c.v	16.052	5.77	6.20
Min rank	1	1	1
Max rank	1775	1775	1321.50
For linear fit y=a+bx *Parameters*	a =3.05 b = -0.96	a =2.99 b = -0.91	a =3.03 b = -0.93
Standard Error	0.057	0.039	0.045
Correlation Coefficient	0.995	0.997	0.997

Table 4.3: *Comparison of different ranking models*

Here Standard Error (S) is the standard error of the estimate which quantifies the spread of data points around the regression curve and Correlation Coefficient (r) is the square-root of the normalized difference between the spread around mean and spread around the fitting function. As the regression model better describes the data, the correlation coefficient will approach unity. It can be seen that the random texts taken from the computer science literature do exhibit Zipf-like distribution with the slope of the linear regression touching unity. However, there is a marked difference in the performance of Maximal Rank and Tied Rank verses Random Rank of Zipf. There is a need to see whether the alternative ranking procedures perform better in other texts.

As far as the distribution of rank and frequency are concerned, it is found that the relation is a Shifted Power distribution (Mandelbrot Zipf's law) of the form

$$g(r) = a(r+b)^c$$

where the coefficients are estimated as a = 3301.44, b = -2.99 and c = -1.23

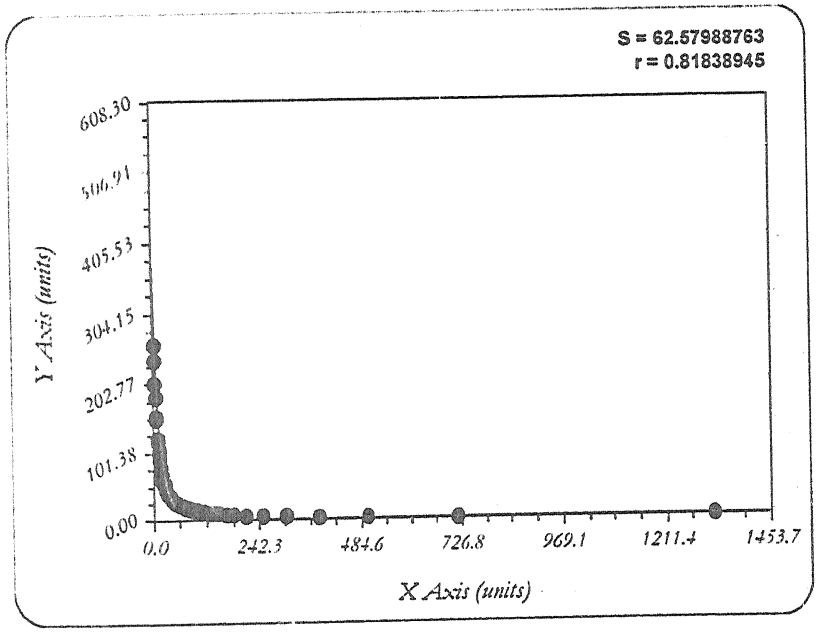


Figure 4.1: *Plot of tied-rank (x-axis) vs. frequency for the random text from computer science literature*

Where S and r are as defined above. It can been seen that the power distribution (Mandelbrot Zipf's law) is fitting this type of data fairly well but with a slight modification in the form and parameters for different texts.. Besides this, the authors

plotted the log rank with log frequency to see how the ranking methods fare. Here, the x-axis refers to the log-rank and y-axis to the log frequency

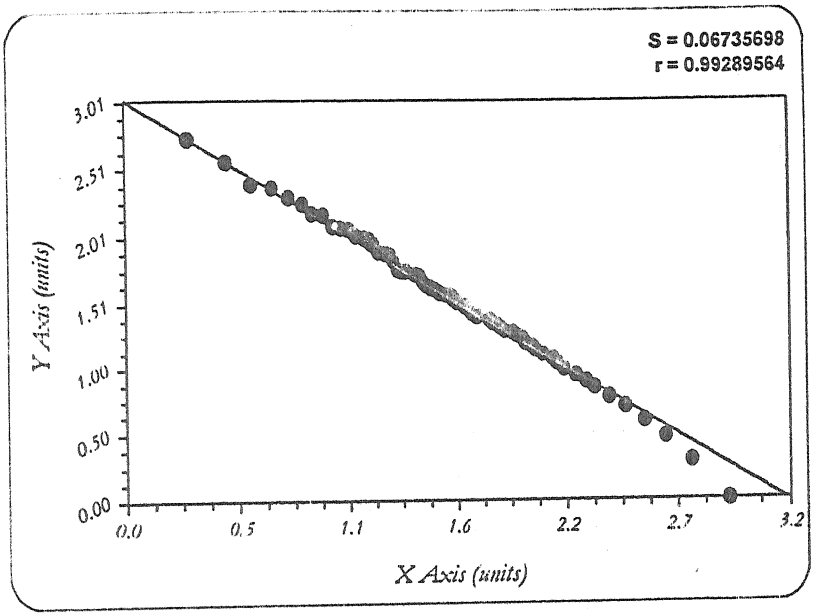


Figure 4.2: *Plot of log rank with log freq. for random rank method for Computer Science Literature*

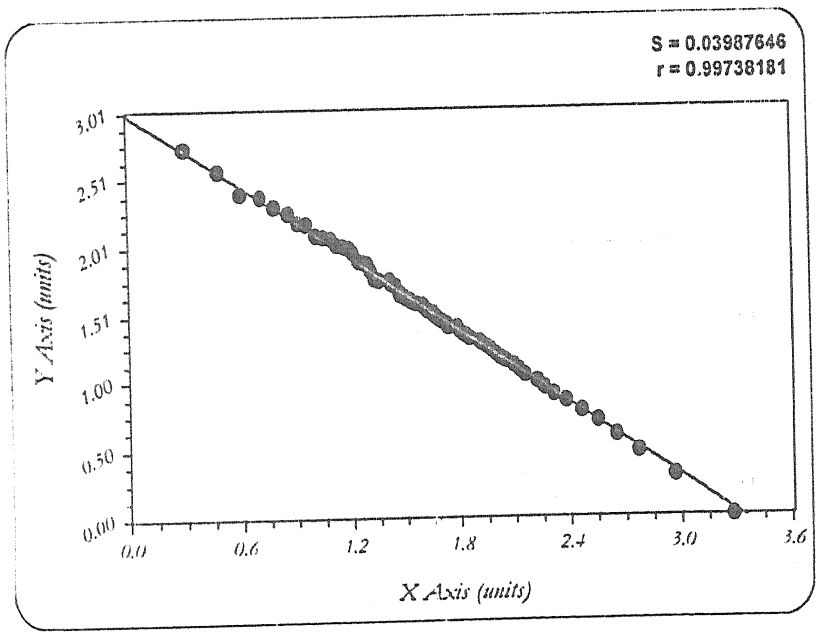


Figure 4.3: *Plot of log rank with log freq. for maximal rank method for Computer Science Literature*

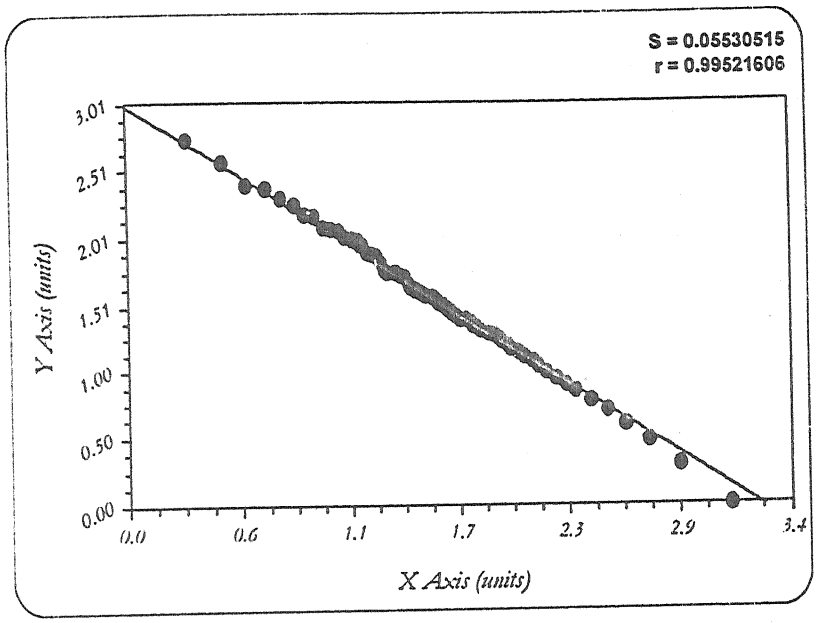


Figure 4.4: *Plot of log rank with log freq. for tied rank method for Computer Science Literature*

It could be seen very clearly that both the Maximal rank method and Tied rank method perform better than the Random rank method of Zipf. It can be seen from the fits of the rank-range at the end.

Discussion

From the figures given in the earlier section it is evident that the lower tail (containing lower ranks) of the plot of log rank vs. log frequency behaved in the best possible manner in the case of Maximal rank. The scatter in tied rank method was better than that in random rank method but not better than that in the maximal rank method. The question that naturally arises is whether the ranking method had a bearing on the type of text in question.

Conclusion

There are two basic issues, which come out of this exercise. Firstly, random texts do follow Zipf's law, however the exponent varies from text to text. The method of random rank performs inferiorly to the maximal rank method and the tied rank method proposed by authors, however there is a need for further investigation in this area as to ascertain whether the ranking method has a bearing on the type of text in question.

rank(ran)	g(r)	r(max)	g(rmax)	r(tied)	g(rt)
1	553	1	553	1	553
2	545	2	545	2	545
3	375	3	375	3	375
4	259	4	259	4	259
5	238	5	238	5	238
6	204	6	204	6	204
7	184	7	184	7	184
8	155	8	155	8	155
9	153	9	153	9	153
10	124	10	124	10	124
11	121	11	121	11	121
12	118	12	118	12	118
13	105	13	105	13	105
14	103	14	103	14	103
15	99	15	99	15	99
16	92	16	92	16	92
17	79	17	79	17	79
18	77	18	77	18	77
19	76	19	76	19	76
20	68	20	68	20	68
21	59	21	59	21	59
22	58	22	58	22	58
23	57	25	57	24	57
26	54	26	54	26	54
27	53	27	53	27	53
28	47	28	47	28	47
29	45	29	45	29	45
30	43	30	43	30	43
31	42	31	42	31	42
32	41	32	41	32	41
33	40	33	40	33	40
34	39	35	39	34.5	39
36	37	38	37	37	37
39	36	39	36	39	36
40	35	40	35	40	35
41	33	41	33	41	33
42	32	44	32	43	32
45	31	45	31	45	31

46	30	46	30	46	30
47	29	47	29	47	29
48	28	48	28	48	28
49	27	51	27	50	27
52	26	52	26	52	26
53	25	59	25	56	25
60	24	60	24	60	24
61	23	63	23	62	23
64	22	66	22	65	22
67	21	69	21	68	21
70	20	77	20	73.5	20
78	19	80	19	79	19
81	18	86	18	83.5	18
87	17	89	17	88	17
90	16	96	16	93	16
97	15	99	15	98	15
100	14	108	14	104	14
109	13	121	13	115	13
122	12	128	12	125	12
129	11	138	11	133.5	11
139	10	158	10	148.5	10
159	9	175	9	167	9
176	8	193	8	184.5	8
194	7	228	7	211	7
229	6	276	6	252.5	6
277	5	338	5	307.5	5
339	4	430	4	384.5	4
431	3	568	3	499.5	3
569	2	867	2	718	2
868	1	1775	1	1321.5	1

Table 4.4: *Ranks & Rank Frequencies by different ranking methods*

Second Analysis: *Whether the distribution of words according to their length and the hits they are able to generate on the popular search engine "Google" follows Zipf's Law.*

Introduction

Any text is made up of words of variable length. The distribution of words contained in a text itself is of great interest to scientists. Web is probably the largest mass of words of various kinds. Many scientists have attempted to examine the informetric properties of the web in the past. Rousseau[12] (2001) has tried to analyze a time series of the number of hits of word "Euro" on the web during a period of one year. Lee Breslau[13] et. al.(1999) raised an issue that whether web requests from a fixed user community are distributed according to Zipf's law. They found that the page request distribution seen by web proxy caching using traces from a variety of sources does not follow Zipf's distribution precisely, but instead follows a Zipf-like distribution with varying exponents. Chao & D'haeseleer[14] (2001) attempted to find the distribution of Variable length Phatic interjectives on the World Wide Web. They found that the number of pages found containing these words would fall off as a power law. However the exponents for length frequency distributions of different interjectives were much larger than –1 predicted by Zipf's law. In this paper, we have tried to examine the distribution of words according to their length and the hits they are able to generate on the popular search engine "Google".

Method and Analysis

We have taken the hits at a particular point of time just to take a rough estimate of the distribution of these words on the Internet. "Google" offers a scale-free network a-priori as it crawls the web from its current database. However it will be good to try the similar search on different search engines. Driven by Bar-Ilan[15] (2001) as cited in Rousseau[12] (2001) that the most cybermetric research results more in statements of principle than in exact results, Hits at a particular point of time were taken just to take a rough estimate of the distribution of these words on the internet. The constraints in getting the data, which stood unchanged for longer period, were accepted and an empirical approach was taken to explore the distribution.

The authors have taken a text from a computer science " Operating System - Concepts and Design", by Milan Milenkovic, Second edition, 1997 (Tata McGraw Hill, New Delhi). The authors have counted the frequency of occurrence of each unique word in the text, and found 1775 unique or different words out of a total of 10,043 words in the full text. We searched for the number of hits each unique word obtained on the search engine "Google".

We tried to find out whether there is any relation between the word length (i.e. the number of alphabets in the word) and the number of hits it gets. The distribution of words has the following descriptive properties

Word-length	Freq	Average log hits	Log length	Minimum	Maximum	Mean	Std. Deviation
1	2	8.87	0.00	8.51	9.23	8.87	0.51
2	21	8.61	0.30	7.49	9.25	8.61	0.38
3	59	7.72	0.48	5.27	9.43	7.72	0.83
4	142	7.58	0.60	5.45	8.76	7.58	0.57
5	177	7.19	0.70	5.35	8.42	7.19	0.61
6	220	6.93	0.78	5.12	8.28	6.93	0.59
7	271	6.85	0.85	4.41	8.11	6.85	0.65
8	246	6.70	0.90	2.53	8.09	6.70	0.77
9	191	6.57	0.95	3.34	7.9	6.57	0.66
10	155	6.39	1.00	3.18	7.84	6.39	0.80
11	117	6.33	1.04	3.96	8.33	6.33	0.69
12	64	6.03	1.08	3.38	7.46	6.04	0.89
13	39	5.82	1.11	3.43	7.32	5.82	1.03
14	19	5.53	1.15	3.08	7.36	5.53	1.23
15	19	5.13	1.18	3.7	6.92	5.13	0.84
16	9	4.91	1.20	3.59	6.54	4.91	0.87
17	7	4.78	1.23	3.6	5.57	4.78	0.72
18	4	4.15	1.26	3.84	4.81	4.15	0.45
19	3	5.65	1.28	5.18	6.09	5.65	0.46
20	1	2.41	1.30	2.41	2.41	2.41	.
21	1	4.21	1.32	4.21	4.21	4.21	.
22	1	5.50	1.34	5.5	5.5	5.50	.
23	1	5.48	1.36	5.48	5.48	5.48	.

Table 4.5: *Descriptive Statistics of the words used in Computer Sc Literature*

The distribution of words in the given text follows a distribution similar to Hoerl Model of the form $y = ab^x x^c$. It can be seen that barring the words with lengths >19, all word-lengths have made appearances more than once. In fact if we do not consider the words

with length more than 16, we may treat this distribution as a Gaussian distribution. This might not reflect anything at this stage but when we plot the Log-length with the log average hits this would have a huge impact on the inferences drawn.

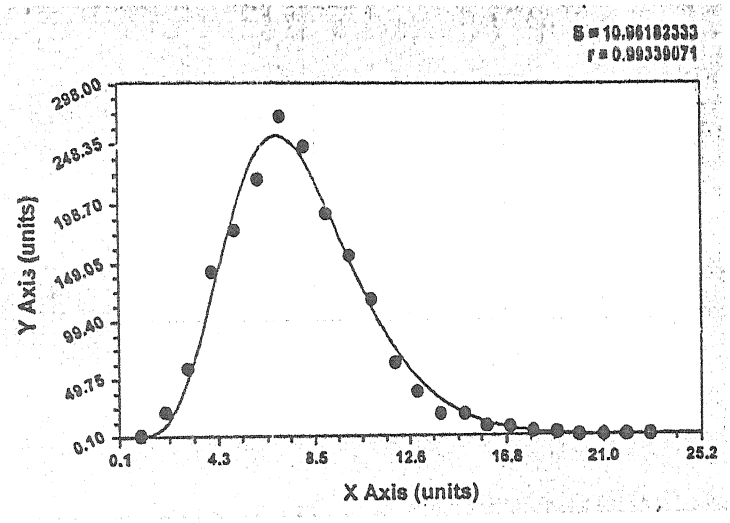


Figure 4.5: *Distribution of words w.r.t. length vs. frequency*

To see whether there is a law embedded in the distribution we plotted the data obtained in the following three manners.

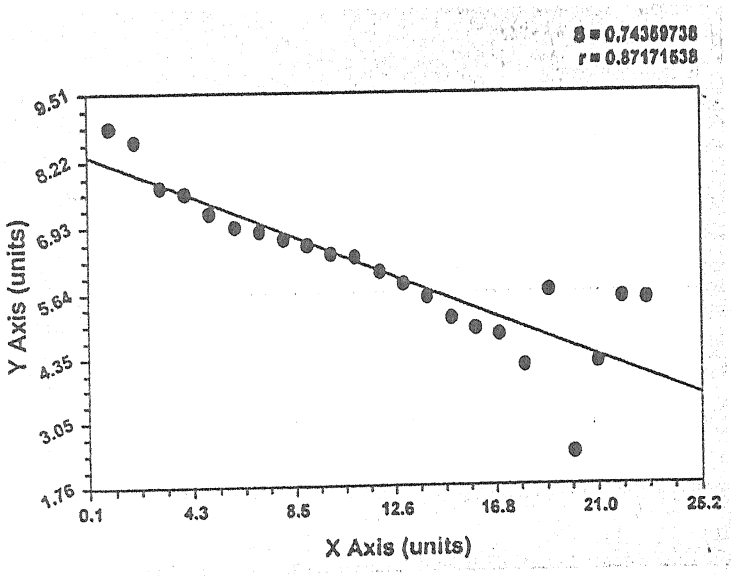


Figure 4.6: *Plot of Length (all) vs. average log hits*

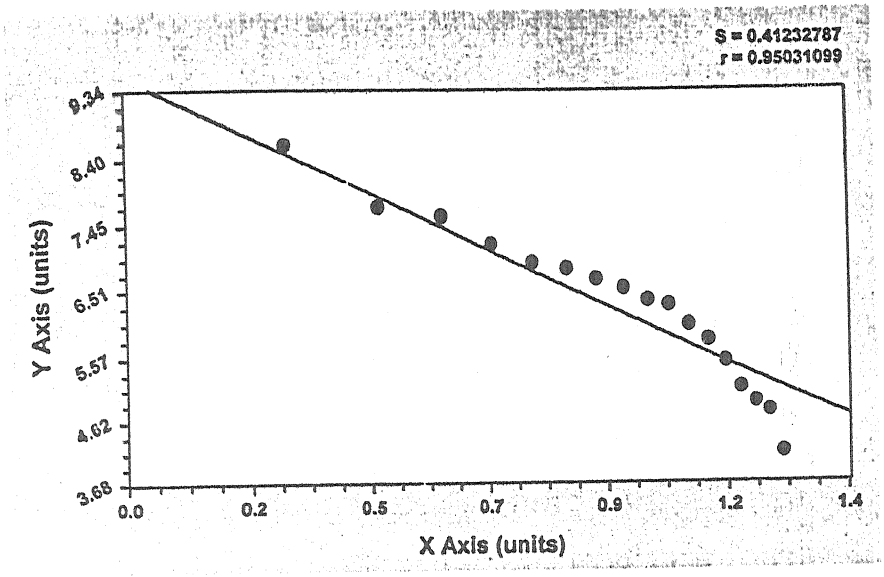


Figure 4.7: *Plot of Log -Length (Up to 18) vs. average log hits*

when the linear model y=a+bx is fitted here, the fit performed as above. The coefficient Data was a =9.53 and b = -3.51. So one can conclude that Zipf's law is weakly applicable here too.

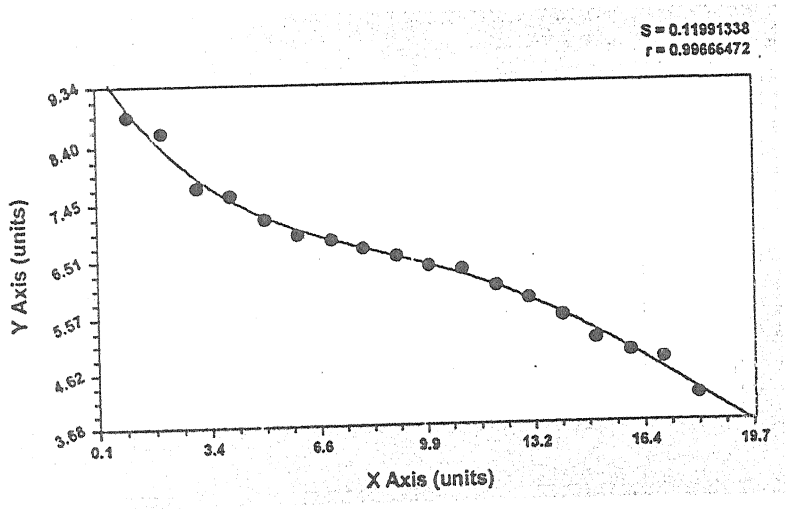


Figure 4.8: *Plot of Length (Up to 18) vs. average log hits*

It can be seen that the data is fairly well described by the 4th Degree Polynomial Fit of the form $y = a + bx + cx^2 + dx^3$ when log length was plotted against log hits. However this result excludes some values i.e. the values for which the word length is up to 18. The

parameters of 4th Degree Polynomial Fit were a =9.79, b = -0.91, c=0.11, d =-0.006 and e =0.0001.

Conclusion

This result found is very similar to one found by Chao & D'haeseleer[14] (2001) for the distribution of Variable length phatic interjectives on the World Wide Web. It is a Zipf type distribution with exponent not close to unity (In fact it came out to be –3.51). However, interesting thing here is that there do exist some relation between length of the word and the number of hits it gets on the web search engine "Google" provided the word length remained less than 18 characters.

Section 2: Zipf's Law in English Literature *(Aladdin and the Wonder Lamp[33])*

For this section, we have selected The Project Gutenberg e-text of "Aladdin and the Wonder Lamp", a "public domain" work distributed by Professor Michael S. Hart through the Project Gutenberg Association. Project Gutenberg is the oldest producer of free ebooks on the Internet (http://www.gutenberg.org/).

The choice of this text was done mainly due to the following reasons:

1. It is a popular work and it is written in a very simple manner so the words used are easy to understand and commonly used.
2. German translation of this text was available so that a comparison can be made between the English and German language after the analysis of the text.

The following statistics were obtained for the text under consideration.

Statistic for the document	
Number of words	5319
Number of sentences	661
Number of words per sentence	8.05
Number of syllables per word (approximate)	1.54
Flesch index[1]	68.23

The Flesch index of readability for this document is 68% this means it is a fairly easy document to understand and is of the level of eighth standard. The prominent words, which were obvious to get more occurrences in this text were, genie (24), lamp (24), mother (25), palace (30), magician (32), sultan (43), princess (45) and Aladdin (98). There were lot of occurrences for the supporting words like, a, of, he, and, to & the.

The zipfian data for the text is obtained and is presented on the table given on the next page.

[1] For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

Rank	Frequency	Log rank	Log Frequency
1	385	0	2.6
2	222	0.3	2.3
3	165	0.5	2.2
4	122	0.6	2.1
5	109	0.7	2
6	98	0.8	2
7	96	0.8	2
8	89	0.9	1.9
9	88	1	1.9
10	87	1	1.9
11	81	1	1.9
12	68	1.1	1.8
13	56	1.1	1.7
14	54	1.1	1.7
15	52	1.2	1.7
16	50	1.2	1.7
17	46	1.2	1.7
18	45	1.3	1.7
20	44	1.3	1.6
21	43	1.3	1.6
23	40	1.4	1.6
25	38	1.4	1.6
26	34	1.4	1.5
27	32	1.4	1.5
29	31	1.5	1.5
30	30	1.5	1.5
32	27	1.5	1.4
33	26	1.5	1.4
35	25	1.5	1.4
38	24	1.6	1.4
42	21	1.6	1.3
45	20	1.7	1.3
47	19	1.7	1.3
48	18	1.7	1.3
53	16	1.7	1.2
54	15	1.7	1.2
59	14	1.8	1.1
60	13	1.8	1.1
63	12	1.8	1.1
66	11	1.8	1
77	10	1.9	1
86	9	1.9	1
97	8	2	0.9
112	7	2	0.8
127	6	2.1	0.8
147	5	2.2	0.7
181	4	2.3	0.6
240	3	2.4	0.5
334	2	2.5	0.3
524	1	2.7	0

Table 4.6: *Zipfian data for the text (Aladdin and the Wonder Lamp)*

Word	Occurence	Rank
a	96	7
aladdin	98	6
of	109	5
he	122	4
to	165	3
and	222	2
the	385	1

Table 4.7: *Most occurring words (Aladdin and the Wonder Lamp)*

Linguists are puzzled by the phenomena that most words are not used much while some occur many a time. Zipf explained this and called it "principle of least effort. He claimed that people minimized their efforts in using language. Zipf's law thus became a feature of human language.

With this framework and preliminary analysis, we proceeded for the regression and curve fitting for this text.

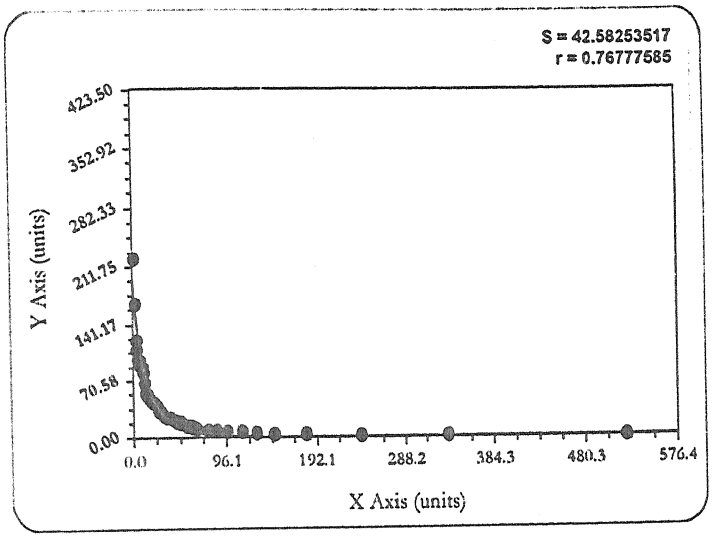


Figure 4.9: *Plot of rank & frequency in Aladdin and the Wonder Lamp*

Bleasdale Model $y = (a + bx)^{\frac{-1}{c}}$ with coefficient data a = 0.15, b = 0.005 and c = 0.36 fitted this data in the manner which is shown in the graph given above. Bleasdale model is a yield-density type models. The prominent characteristic of this model is that if the response is such that as density (x) increases, but the yield (y) approaches a fixed value, the relationship is asymptotic.

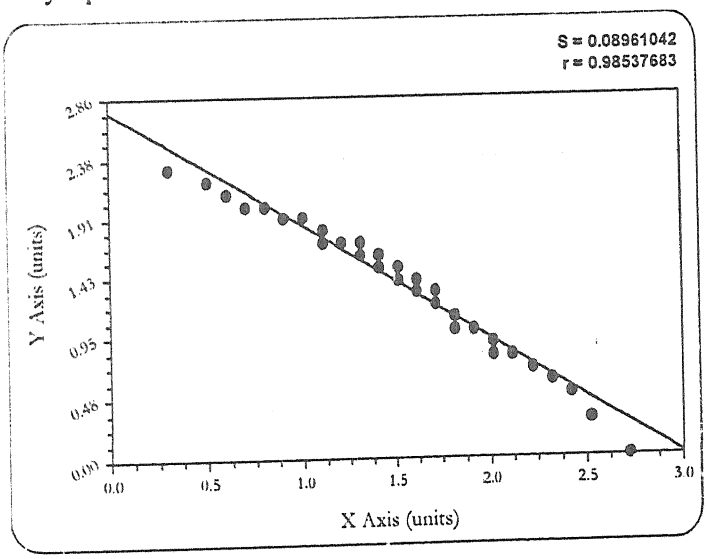


Figure 4.10: *Plot of log rank & log frequency in Aladdin and the Wonder Lamp*

The Linear Fit y = a+bx for the log rank and log frequency data obtained coefficients as a = 2.76 and b =-0.92. The residual[2] plot, which shows the difference between the data points and the model, evaluated at the data points is shown below.

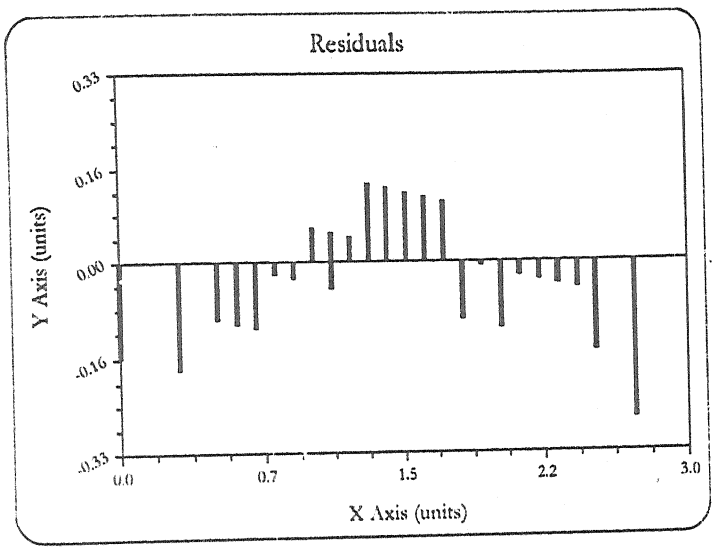


Figure 4.11: *Residual Plot for data points & model in Aladdin and the Wonder Lamp*

The result verified that Zipf's law is applicable in this text and for the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$) the coefficient c in this case is –0.92.

[2] The residual at point k is defined by $\text{Residual}_k = y_k - f(x_k)$

Where y_k is the measured value at x_k, and $f(x_k)$ is the predicted value at x_k.

Section 3: Zipf's Law in German Literature *(Aladdin und die Wunderlampe[34])*

For this section, we have selected The Project Gutenberg e-text of "Aladdin und die Wunderlampe", by Ludwig Fulda with original illustration by Max Liebert. Project Gutenberg is the oldest producer of free e-books on the Internet (http://www.gutenberg.org/).

The choice of this text was done mainly due to the fact that it is a popular work. English translation of this text was available so that a comparison can be made between the English and German language after the analysis of the text. However this version is a more elaborated one as it contains illustrations also. This can be verified by the fact that it contains almost three times of the words in the English version.

The following statistics were obtained for the text under consideration.

Statistic for the document	
Number of words	17686
Number of sentences	3536
Number of words per sentence	5.00
Number of syllables per word (approximate)	1.70
Flesch index[3]	57.90

The Flesch index of readability for this document is around 58% this means it is a fairly easy document to understand and is of the level of high school standard. The prominent words, which were obvious to get more occurrences in this text were, "und" (and), "die" (the), "er" (he), "zu" (to), "in" (in), "sich" (itself) and "von" (from). So it is reaffirmed that there were lot of occurrences for the supporting words like, a, of, he, and, to & the. In the English version we found the most happening words, which are context specific and are typical for the subject of the text. We tried to make a comparison of these words in English and German versions. The table given below gives an account of the

[3] For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

frequency of occurrence of these words in the texts. Since the number of words differs in both the texts we have shown the % of frequency of occurrence.

English word	Frequency	German word	Frequency
genie (24)	24	Flaschengeist	0*
lamp (24)	24	Lampe	30
mother (25)	25	Mutter	53
palace (30)	30	Palast	16
magician (32)	32	Zauberer	2*
sultan (43)	43	Sultan	61
princess (45)	45	Prinzessin	27
Aladdin (98)	98	Aladdin	88

**- These words may be used in a different form in German text.*

Table 4.8: *English Words vs German Words & their frequency in Aladdin und die Wunderlampe*

There were 4929 different words found in this analysis. The zipfian data for the text is obtained and is presented on the table given on the next page.

Rank	Freq	Log Rank	Log Freq
1	561	0.00	2.75
2	426	0.30	2.63
3	371	0.48	2.57
4	365	0.60	2.56
5	262	0.70	2.42
6	219	0.78	2.34
7	209	0.85	2.32
8	200	0.90	2.30
9	193	0.95	2.29
10	187	1.00	2.27
11	177	1.04	2.25
12	164	1.08	2.21
13	160	1.11	2.20
14	153	1.15	2.18
15	147	1.18	2.17
16	132	1.20	2.12
17	126	1.23	2.10
18	125	1.26	2.10
19	117	1.28	2.07
20	112	1.30	2.05
21	109	1.32	2.04
22	104	1.34	2.02
23	99	1.36	2.00
24	96	1.38	1.98
25	95	1.40	1.98
27	92	1.43	1.96
28	91	1.45	1.96
29	90	1.46	1.95
30	88	1.48	1.94
31	85	1.49	1.93
32	83	1.51	1.92
34	76	1.53	1.88
35	75	1.54	1.88
36	74	1.56	1.87
37	64	1.57	1.81
38	62	1.58	1.79
42	61	1.62	1.79
43	59	1.63	1.77
44	57	1.64	1.76
46	55	1.66	1.74
47	54	1.67	1.73
49	53	1.69	1.72
50	50	1.70	1.70
52	49	1.72	1.69
53	47	1.72	1.67
54	46	1.73	1.66
56	45	1.75	1.65
57	43	1.76	1.63
61	42	1.79	1.62

62	41	1.79	1.61
63	40	1.80	1.60
64	37	1.81	1.57
66	36	1.82	1.56
67	35	1.83	1.54
70	34	1.85	1.53
71	33	1.85	1.52
72	32	1.86	1.51
74	31	1.87	1.49
77	30	1.89	1.48
81	29	1.91	1.46
83	28	1.92	1.45
85	27	1.93	1.43
86	26	1.93	1.41
90	25	1.95	1.40
94	24	1.97	1.38
100	23	2.00	1.36
103	22	2.01	1.34
104	21	2.02	1.32
108	20	2.03	1.30
110	19	2.04	1.28
117	18	2.07	1.26
118	17	2.07	1.23
131	16	2.12	1.20
137	15	2.14	1.18
144	14	2.16	1.15
151	13	2.18	1.11
164	12	2.21	1.08
176	11	2.25	1.04
192	10	2.28	1.00
218	9	2.34	0.95
238	8	2.38	0.90
275	7	2.44	0.85
319	6	2.50	0.78
361	5	2.56	0.70
453	4	2.66	0.60
602	3	2.78	0.48
911	2	2.96	0.30
1633	1	3.21	0.00

Table 4.9: *Zipfian data for the text (Aladdin und die Wunderlampe)*

Linguists are puzzled by the phenomena that most words are not used much while some occur many a time. Zipf explained this and called it "principle of least effort. He claimed that people minimized their efforts in using language. Zipf's law thus became a feature of human language. With this framework and preliminary analysis, we proceeded for the regression and curve fitting for this text.

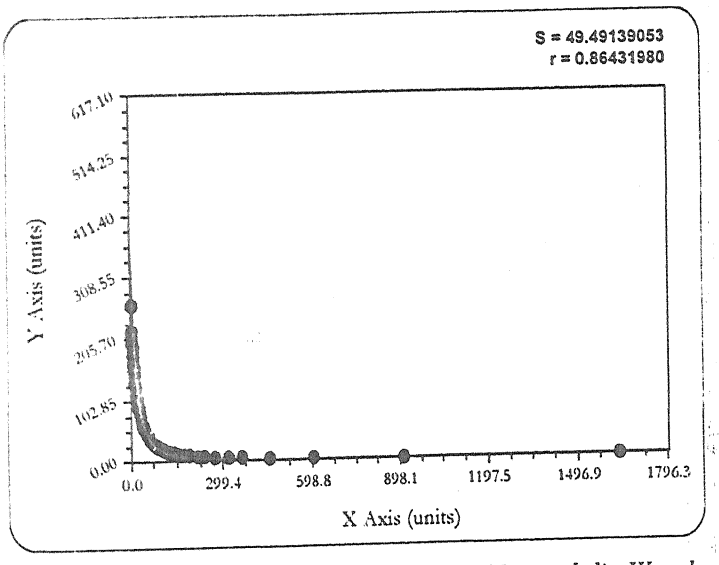


Figure 4.12: *Plot of ranks vs. Frequency in Aladdin und die Wunderlampe*

Bleasdale Model $y = (a + bx)^{\frac{-1}{c}}$ with coefficient data a = 0.06, b = 0.002 and c = 0.47 fitted this data in the manner which is shown in the graph given above. Bleasdale model is a yield-density type models. The prominent characteristic of this model is that if the response is such that as density (x) increases, but the yield (y) approaches a fixed value, the relationship is asymptotic.

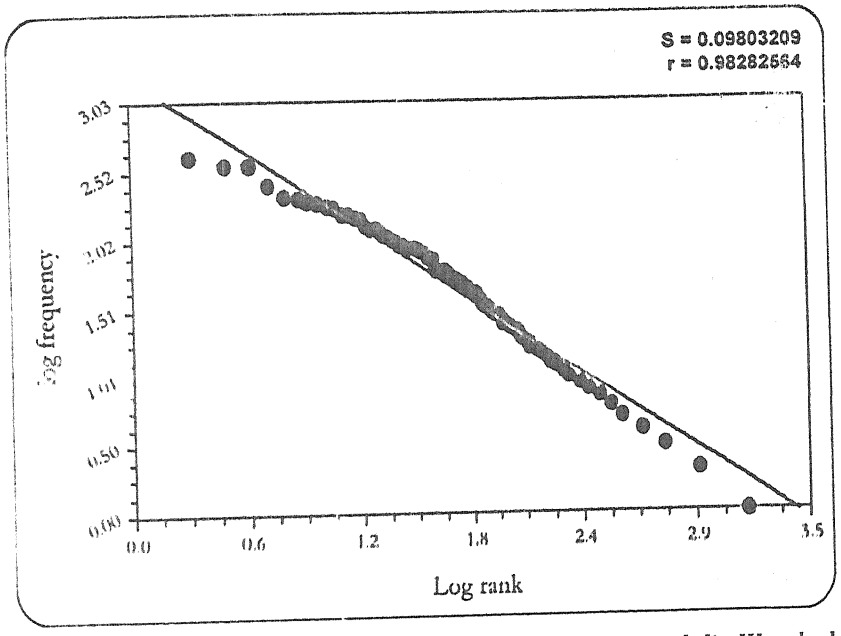


Figure 4.13: *Plot of log ranks vs. log frequency in Aladdin und die Wunderlampe*

The Linear Fit y = a+bx for the log rank and log frequency data obtained coefficients as a = 3.20 and b =-0.92. The residual plot, which shows the difference between the data points and the model, evaluated at the data points is shown below.

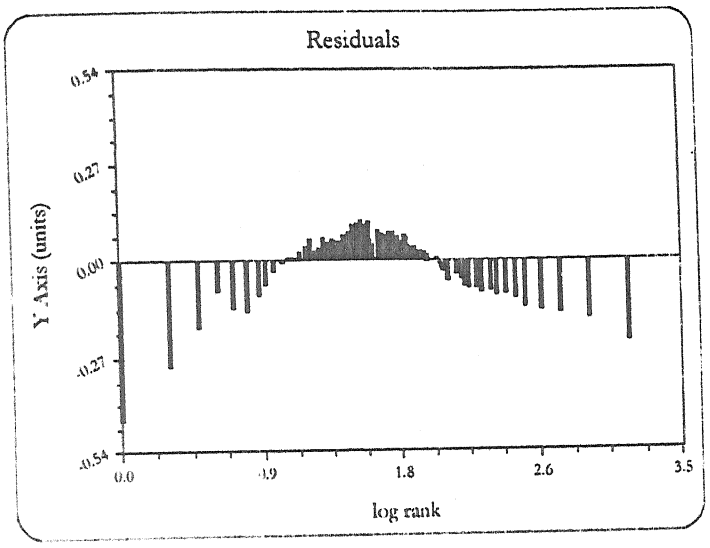


Figure 4.14: *Residual Plot for data points & model in in Aladdin und die Wunderlampe*

The result verified that Zipf's law is applicable in this text and for the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$) the coefficient c in this case is –0.92.

Section 4: Zipf's Law for English-German Business Dictionary *(Mr. Honey's Small Business Dictionary (English-German)* [35]

For this section, the Project Gutenberg E-text of "Mr. Honey's Small Business Dictionary (English-German)" by Winfried Honig was taken up. Mr. Honey (Winfried Honig) compiled English/German dictionaries for almost 3 decades to provide his colleagues and students with samples of the language of business and highlight the need for special dictionaries covering the special language used in different branches of the industry. These wordlists are now fed into the LEO Online Dictionary (http://dict.leo.org) and the DicData Online Dictionary (http://www.dicdata.de).

The choice of this text was done mainly due to the following reasons:

- Dictionary does not follow any linguistic style of writing. They only depict the words or group of words whose meaning are to be given.
- Small Business dictionary was taken as it was specifically devoted to business words.
- Translation from German to English was given. The benefit derived from this lies in the opportunity we got in investigating the number of words required in the other language to explain the original word.

We have separated the English words from the German words and made two text files. The following statistics were obtained for the English part under consideration.

Statistic for the document	
Number of words	10089
Number of sentences	5763
Number of words per sentence	1.75
Number of syllables per word (approximate)	2.54
Flesch index[4]	-9.95

The Flesch index of readability for this document is around -10%. This means it is a fairly difficult document to understand and is of the level of a law school graduate. The

[4] For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

prominent words, which were obvious to get more occurrences in this text were, particles like "of", "a", "and", "on" and "for" & the typical words for this type of text like "price", "goods", "account", "tax", "capital", "market", "trade" and "costs". These words also appear more number of times because dictionary was containing a combination of these words like "abandon a business", "abandon a plan", and "abandon a project" were the three entries in the dictionary so "abandon" has come three times and also "plan", "business" and "project" got one more count. The same would be true for the German part.

There were 372 different words found in this analysis. The zipfian data for the text is obtained and is presented on the table given on the next page.

Rank	Freq	Log Rank	Log Freq
1	580	0.00	2.76
2	95	0.30	1.98
3	60	0.48	1.78
4	57	0.60	1.76
5	51	0.70	1.71
6	50	0.78	1.70
7	47	0.85	1.67
10	42	1.00	1.62
13	40	1.11	1.60
14	38	1.15	1.58
15	35	1.18	1.54
17	34	1.23	1.53
19	33	1.28	1.52
23	32	1.36	1.51
24	31	1.38	1.49
25	30	1.40	1.48
26	29	1.41	1.46
28	28	1.45	1.45
32	27	1.51	1.43
35	26	1.54	1.41
36	25	1.56	1.40
37	24	1.57	1.38
39	23	1.59	1.36
42	22	1.62	1.34
44	21	1.64	1.32
46	20	1.66	1.30
49	19	1.69	1.28
52	18	1.72	1.26
58	17	1.76	1.23
65	16	1.81	1.20
67	15	1.83	1.18
75	14	1.88	1.15
86	13	1.93	1.11
95	12	1.98	1.08
117	11	2.07	1.04
133	10	2.12	1.00
150	9	2.18	0.95
174	8	2.24	0.90
205	7	2.31	0.85
261	6	2.42	0.78
335	5	2.53	0.70
422	4	2.63	0.60
586	3	2.77	0.48
858	2	2.93	0.30
1402	1	3.15	0.00

Table 4.10: *Zipfian data for the text (English part) for Mr. Honey's Small Business Dictionary (English-German)*

Zipf "principle of least effort" is a not a valid connotation here as dictionary mentions all the words that are required for the small business. It will thus be important to see whether the zipf's law is applicable in this context. The hypothesis that people minimized their efforts in using language is also not applicable here.

With this framework and preliminary analysis, we proceeded for the regression and curve fitting for this text.

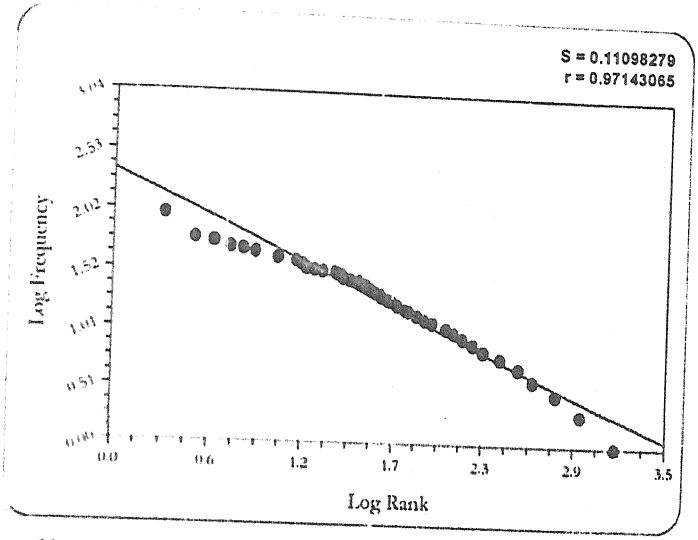


Figure 4.15: *Plot of log ranks vs. log frequency in for Mr. Honey's Small Business Dictionary*

The Linear Fit y = a+bx for the log rank and log frequency data obtained coefficients as a = 2.35 and b =-0.66. The residual plot, which shows the difference between the data points and the model, evaluated at the data points is shown below.

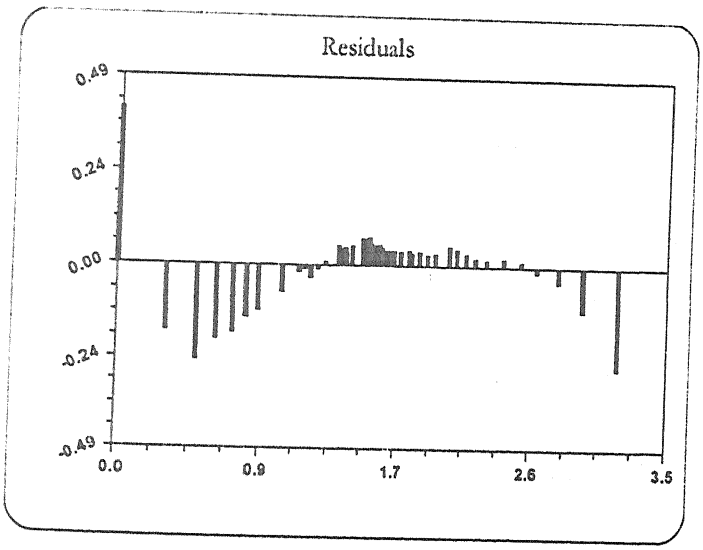


Figure 4.16: *Residual Plot for data points & model in Mr. Honey's Small Business Dictionary*

The result verified that Zipf's law is not applicable in this text and for the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$) the coefficient c in this case is –0.66 that is not close to –1.

The following statistics were obtained for the German part under consideration.

Statistic for the document	
Number of words	9107
Number of sentences	5792
Number of words per sentence	1.57
Number of syllables per word (approximate)	3.41
Flesch index[5]	-82.83

The Flesch index of readability for this document is around -82%. This means it is an extremely difficult document to understand and is of the level of a law school graduate. The prominent words, which were obvious to get more occurrences in this text were, particles like "of", "a", "and", "on" and "for" which are shown in the following table. The typical words for this type of text are also given in the third column of the following table:

German	English	Frequency	German	English	Frequency
der	the	94	einer	one	24
in	in	42	ein	one	22
auf	up	41	preis	price	21
us	us	40	ware	commodity	20
des	of	34	einen	one	18
von	the	34	kosten	cost	18
nicht	not	32	gesetz	law	16
eines	one	30	markt	market	16
sich	itself	30	nachfrage	inquire	16
und	and	30	angebot	offer	15

Table 4.11: *German Words, their meaning & frequency in Mr. Honey's Small Business Dictionary*

[5] For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

The meaning of "eines", "einen", "einer:, "ein" is "one" but different form is used at different places. This might be a special property of the German language. There were 5913 different words found in this analysis. The zipfian data for the text is obtained and is presented on the table given on the next page.

Rank	Freq	Log Rank	Log Freq
1	94	0.00	1.97
2	43	0.30	1.63
3	42	0.48	1.62
4	41	0.60	1.61
5	40	0.70	1.60
6	34	0.78	1.53
8	32	0.90	1.51
9	30	0.95	1.48
12	25	1.08	1.40
13	24	1.11	1.38
14	23	1.15	1.36
15	22	1.18	1.34
18	21	1.26	1.32
20	20	1.30	1.30
21	18	1.32	1.26
23	16	1.36	1.20
26	15	1.41	1.18
29	14	1.46	1.15
33	13	1.52	1.11
36	12	1.56	1.08
38	11	1.58	1.04
43	10	1.63	1.00
49	9	1.69	0.95
58	8	1.76	0.90
68	7	1.83	0.85
87	6	1.94	0.78
108	5	2.03	0.70
154	4	2.19	0.60
259	3	2.41	0.48
506	2	2.70	0.30
1376	1	3.14	0.00

Table 4.12: *Zipfian data for the text (German) in Mr. Honey's Small Business Dictionary*

Zipf "principle of least effort" is a not a valid connotation here as dictionary mentions all the words that are required for the small business. It will thus be important to see whether the zipf's law is applicable in this context. The hypothesis that people minimized their efforts in using language is also not applicable here. With this framework and preliminary analysis, we proceeded for the regression and curve fitting for this text.

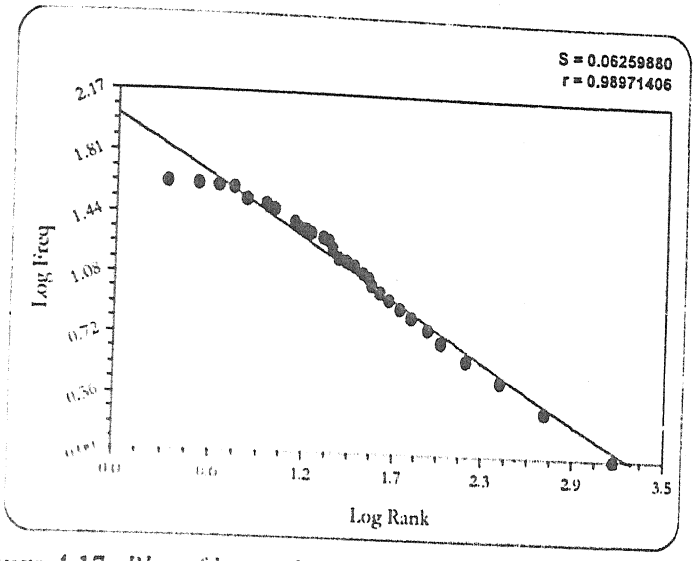


Figure 4.17: *Plot of log ranks vs. log frequency in for Mr. Honey's Small Business Dictionary (German Words)*

The Linear Fit y = a+bx for the log rank and log frequency data obtained coefficients as a = 2.02 and b =-0.63. The residual plot, which shows the difference between the data points and the model, evaluated at the data points is shown below.

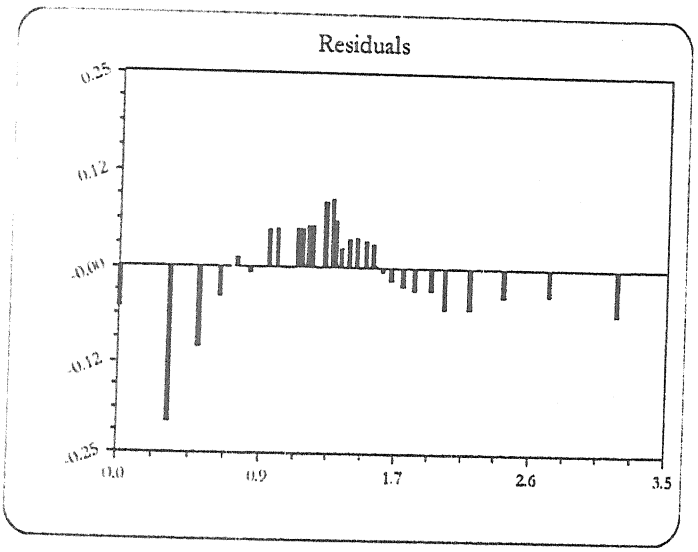


Figure 4.18: *Residual Plot for data points & model in Mr. Honey's Small Business Dictionary (German Words)*

The result verified that Zipf's law is not applicable in this text and for the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$) the coefficient c in this case is –0.66 that is not close to –1.

Section 5: Zipf's Law in Hindi Literature *(Eidgaah by Munshi Premchand[36])*

For this section, we have selected IIT Kanpur's e-text of roman version of a story called "Eidgaah" by Munshi Premchand (http://www.munsipremchand.iitk.ac.in/auther.html). This website has been built as part of a larger effort to create a series of websites based on Indian philosophical texts. This website has been built under a project in the Department of Computer Science & Engineering at the Indian Institute of Technology Kanpur.

The choice of this text was done mainly due to the following reasons:

- It is an epic work by one of the finest writers of Hindi Prose and it is written in a very simple manner so the words used are easy to understand and commonly used.
- The roman version of the original work in Hindi was available by the pioneering work done at IIT Kanpur. It was the part of a website which offers stories of Munshi Premchand in portable document format version (.pdf) with a facility to translate it in many Indian languages. The roman text was obtained in this manner only.
- Dynamic Fonts were used to display Indian languages. But however this did not work for us and thus we downloaded font for roman (DV1-TTYogesh). These are the fonts that are made by Centre for Development of Advanced Computing (CDAC)

The following statistics were obtained for the text under consideration.

Statistic for the document	
Number of words	4951
Number of sentences	505
Number of words per sentence	9.80
Number of syllables per word (approximate)	1.99
Flesch index	28.74

The Flesch index of readability for this document is 28.7% this means it is a document aimed at the level of a college graduate. This was departure from our present understanding that this is a very simple document which is prescribed at the school level books as essential reading and more so as a chapter in some text books. The reason for

this document obtaining a low index may lie in the fact that it is a translated version and that may have changed the structure of sentences. There were lot of occurrences for the supporting words like, "hai", "aura", "hain", "ki", "ke", "men", "to" and "se". The prominent words, which got more occurrences in this text, are as follows.

Word	Frequency
Hameid	69
Nahin	56
Mohasina	31
Paise	28
Lekina	22
Mahamuda	22
Chimata	19
Khilaune	19

Table 4.13: *Most occurring words in Eidgaah*

The Zipfian data for the text is obtained and is presented on the table given below.

Rank	Frequency	Log Rank	Log Freq
1	178	0.00	2.25
2	103	0.30	2.01
3	96	0.48	1.98
4	88	0.60	1.94
5	84	0.70	1.92
6	80	0.78	1.90
8	70	0.90	1.85
9	69	0.95	1.84
10	65	1.00	1.81
11	63	1.04	1.80
12	56	1.08	1.75
13	53	1.11	1.72
14	52	1.15	1.72
15	48	1.18	1.68
16	47	1.20	1.67
17	43	1.23	1.63
18	39	1.26	1.59
19	37	1.28	1.57
20	32	1.30	1.51
21	31	1.32	1.49
23	29	1.36	1.46
24	28	1.38	1.45
25	27	1.40	1.43
26	24	1.41	1.38
27	23	1.43	1.36
29	22	1.46	1.34
31	21	1.49	1.32
32	19	1.51	1.28
34	18	1.53	1.26
36	17	1.56	1.23
38	16	1.58	1.20
42	15	1.62	1.18
45	14	1.65	1.15
51	13	1.71	1.11
61	12	1.79	1.08
65	11	1.81	1.04
75	10	1.88	1.00
84	9	1.92	0.95
94	8	1.97	0.90
109	7	2.04	0.85
127	6	2.10	0.78
151	5	2.18	0.70
193	4	2.29	0.60
257	3	2.41	0.48
364	2	2.56	0.30
571	1	2.76	0.00

Table 4.14: *Zipfian data for the text (Eidgaah)*

With this framework and preliminary analysis, we proceeded for the regression and curve fitting for this text.

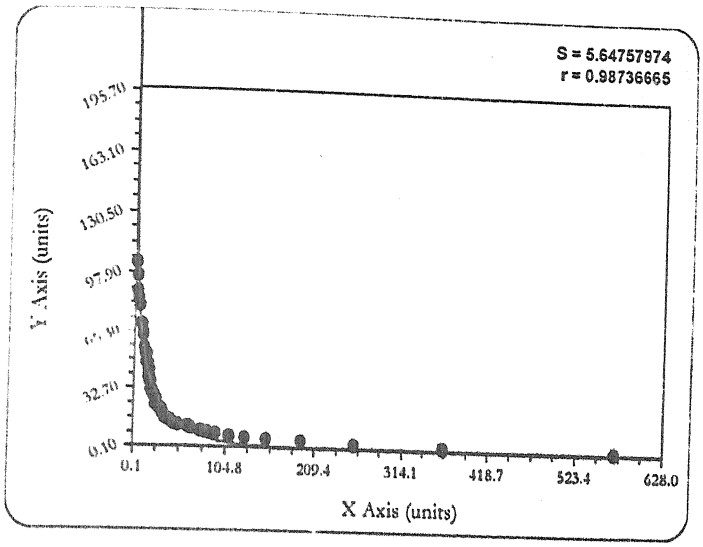


Figure 4.19: *Plot of ranks vs. Frequency in Eidgaah*

Hoerl Model: $y = ab^x x^c$ with coefficient data a = 168.59, b = 0.98 and c = - 0.37 fits this data

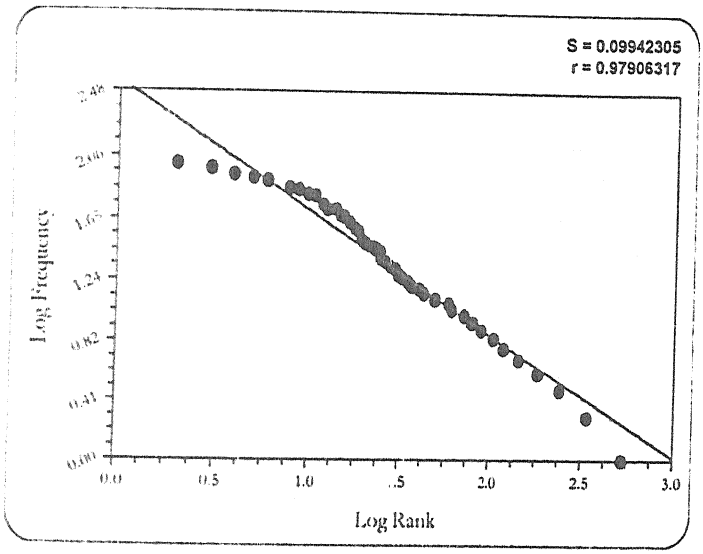


Figure 4.20: *Plot of log ranks vs. log frequency in Eidgaah*

The Linear Fit y = a+bx for the log rank and log frequency data obtained coefficients as a = 2.54 and b = - 0.82. The residual[6] plot, which shows the difference between the data points and the model, evaluated at the data points is shown below.

[6] The residual at point k is defined by $\text{Residual}_k = y_k - f(x_k)$

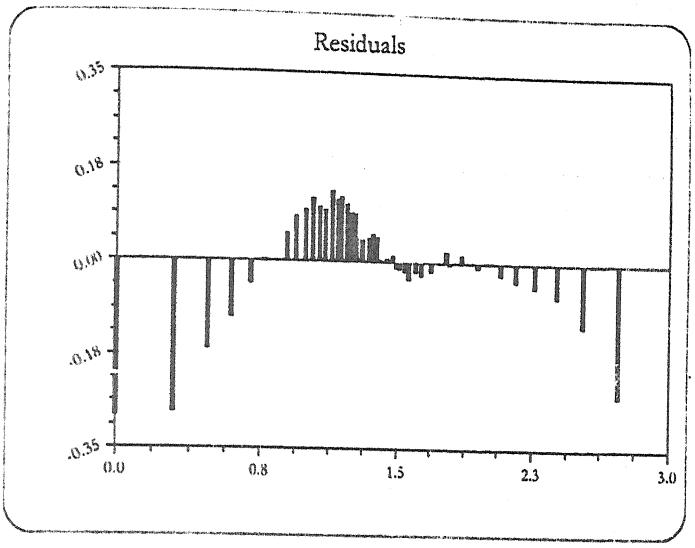


Figure 4.21: *Residual Plot for data points & model in Eidgaah*

The result verified that Zipf's law is applicable in this text and for the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$) the coefficient c in this case is –0.82.

Where y_k is the measured value at x_k, and $f(x_k)$ is the predicted value at x_k.

Section 6: Zipf's Law in a text from Library Science Literature *("The Library[37]", by Andrew Lang)*

For this chapter, we have selected The Project Gutenberg e-text of "The Library", by Andrew Lang #20 in our series by Andrew Lang, December 1999. The choice of this text was done mainly due to the following reasons:

- It is a subject specific work. The aim of selecting this text was to find the pattern of word usage particularly from a text from a subject area.
- It will provide a comparison with the earlier analysis of Computer Science literature i.e. it will enable us to make a comparison between the two subject on the count of applicability of Zipf's law.

The following statistics were obtained for the text under consideration.

Statistic for the document	
Number of words	37498
Number of sentences	5037
Number of words per sentence	7.44
Number of syllables per word (approximate)	1.73
Flesch index[7]	53.33

The Flesch index of readability for this document is 53.33% this means it is a document aimed at the level of a high school student. So this is a very simple document, which can be taken as an elementary reading in the Library science literature. One particular characterstic of this document was the occurrence of lot of alphanumeric words and numbers. Some of these numbers were years while some were page numbers given in the reference. There were 168 different numbers that appeared in this text and they were removed from the Zipf's analysis. The numbers like, 0, 1, 10, 100, 1000, 11, 12, 120, 131

[7] For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

and 13 appeared at various places in the text. A description of most occurring such words is given in the table on the next page.

The prominent numbers, which got more occurrences in this text, are as follows:

Word	Frequency
1	10
2	9
4	9
3	8
1830	6
5	5

Word	Frequency
6	5
10	5
1880	5
7	4
8	4
13	4

Table 4.15: *Numbers & their frequency in "The Library"*

Since this was a text taken from a specific subject area there were bound to be use of context-specific words. The following is an example taken from a words "bibliography" and the its related words. The frequency of occurrence of these words is also shown in the table given below.

Word	Meaning	Frequency
Bibliographical	*Relating to or dealing with bibliography*	2
Bibliography	*A list of writings with time and place of publication (such as the writings of a single author or the works referred to in preparing a document etc.)*	7
Biblioklept	*A person who has a compulsion to steal books*	11
Bibliokleptomaniac	*One who has a morbid tendency to steal books*	1
Biblioklepts	*Persons who has a compulsion to steal books*	4
Bibliomania	*Preoccupation with the acquisition and possession of books*	2
Bibliomaniac	*Person who has a preoccupation with the acquisition and possession of books*	1
Bibliopegia	*Relating to the binding of books*	1
Bibliophile	*Someone who loves (and usually collects) books*	22
Bibliophiles	*Person who loves (and usually collects) books*	7
Bibliotheca	*A collection of books*	1
Bibliothec	*A professional person trained in library science and engaged in library services [syn: librarian]*	1

Table 4.16: *Word meaning & their frequency in "The Library".*
Source: WordNet ® 2.0, © 2003 Princeton University at
http://dictionary.reference.com/browse/Bibliographical

Another striking characteristic found in this document is on the use of connecting and supporting words. We have analyzed such words and found that among top 100 most

occurring words these words are almost capturing 50% of the total words in the document. The table given below shows the frequency of occurrence of such words:

Connecting & Supporting Words	Frequency
The	2746
Of	1913
And	1156
A	906
In	794
To	789
Is	540
His	367
That	314
S	313
He	308
Are	301
For	300
It	299
With	279
As	276
But	243
Be	237
Which	230
Was	225
Or	224
By	220

Connecting & Supporting Words	Frequency
On	205
Not	203
At	186
This	171
Have	160
An	136
Has	136
One	121
Who	121
From	116
There	114
These	114
May	109
All	106
Their	106
De	103
Old	103
They	103
Were	103
So	92
Some	91
More	89
Mr.	89
Like	87

Connecting & Supporting Words	Frequency
Been	86
We	83
When	82
I	79
Had	78
M	76
Them	75
Will	74
Very	71
Most	67
Its	66
No	65
Than	65
If	59
Many	57
First	54
Would	54
Can	53
Such	53
Him	52
Even	51
Our	51
Only	48
Other	48
Total	**17491**

Table 4.17: *Most occurring words (The Library)*

However other subject specific words like "library", "little", "printed", "years", "illustrated", "volumes", "work", "amateur", "edition", "volume", "collection", "English", " modern", "art" and "century" also occurred significantly.

Out of the 6721 distinct words, the zipfian data for the text is obtained and is presented on the table given below.

Rank	Freq	Log Rank	Log Freq
1	2746	0.00	3.44
2	1913	0.30	3.28
3	1156	0.48	3.06
4	906	0.60	2.96
5	794	0.70	2.90
6	789	0.78	2.90
7	540	0.85	2.73
8	367	0.90	2.56
9	314	0.95	2.50
10	313	1.00	2.50
11	308	1.04	2.49
12	303	1.08	2.48
13	301	1.11	2.48
14	300	1.15	2.48
15	299	1.18	2.48
16	279	1.20	2.45
17	276	1.23	2.44
18	243	1.26	2.39
19	237	1.28	2.37
20	230	1.30	2.36
21	225	1.32	2.35
22	224	1.34	2.35
23	220	1.36	2.34
24	206	1.38	2.31
25	205	1.40	2.31
26	203	1.41	2.31
27	186	1.43	2.27
28	171	1.45	2.23
29	160	1.46	2.20
30	136	1.48	2.13
32	121	1.51	2.08
34	116	1.53	2.06
35	114	1.54	2.06
37	109	1.57	2.04
38	106	1.58	2.03
40	103	1.60	2.01
44	92	1.64	1.96
45	91	1.65	1.96
46	89	1.66	1.95
48	87	1.68	1.94
49	86	1.69	1.93
50	83	1.70	1.92
51	82	1.71	1.91
52	79	1.72	1.90
53	78	1.72	1.89
54	76	1.73	1.88
55	75	1.74	1.88
56	74	1.75	1.87
57	71	1.76	1.85
58	67	1.76	1.83
59	66	1.77	1.82
60	65	1.78	1.81
62	59	1.79	1.77
63	57	1.80	1.76
64	55	1.81	1.74
65	54	1.81	1.73
67	53	1.83	1.72
70	52	1.85	1.72
71	51	1.85	1.71
73	48	1.86	1.68
75	47	1.88	1.67
76	46	1.88	1.66
78	45	1.89	1.65
81	44	1.91	1.64
82	43	1.91	1.63
86	42	1.93	1.62
89	41	1.95	1.61
90	40	1.95	1.60
91	39	1.96	1.59
93	38	1.97	1.58
96	37	1.98	1.57
100	36	2.00	1.56
105	35	2.02	1.54
111	34	2.05	1.53
114	33	2.06	1.52
121	32	2.08	1.51
127	31	2.10	1.49
134	30	2.13	1.48
139	29	2.14	1.46
144	28	2.16	1.45
149	27	2.17	1.43
152	26	2.18	1.41
162	25	2.21	1.40
165	24	2.22	1.38
172	23	2.24	1.36
178	22	2.25	1.34
188	21	2.27	1.32
197	20	2.29	1.30
208	19	2.32	1.28
219	18	2.34	1.26
228	17	2.36	1.23
245	16	2.39	1.20
265	15	2.42	1.18
284	14	2.45	1.15
310	13	2.49	1.11
337	12	2.53	1.08
368	11	2.57	1.04
402	10	2.60	1.00
441	9	2.64	0.95
501	8	2.70	0.90
571	7	2.76	0.85
666	6	2.82	0.78
795	5	2.90	0.70
973	4	2.99	0.60
1251	3	3.10	0.48
1750	2	3.24	0.30
2809	1	3.45	0.00

Table 4.18: *Zipfian data for the text (The Library)*

With this framework and preliminary analysis, we proceeded for the regression and curve fitting for this text.

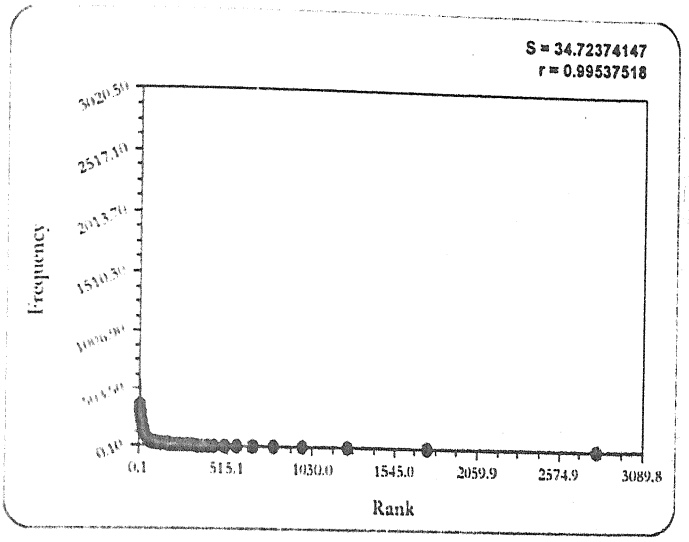


Figure 4.22: *Plot of ranks vs. Frequency in "The Library"*

Modified Hoerl Model, $y = ab^{\frac{1}{x}}x^{c}$ with coefficient data a = 4769.28, b = 0.58 and c = - 1.05 fits this data

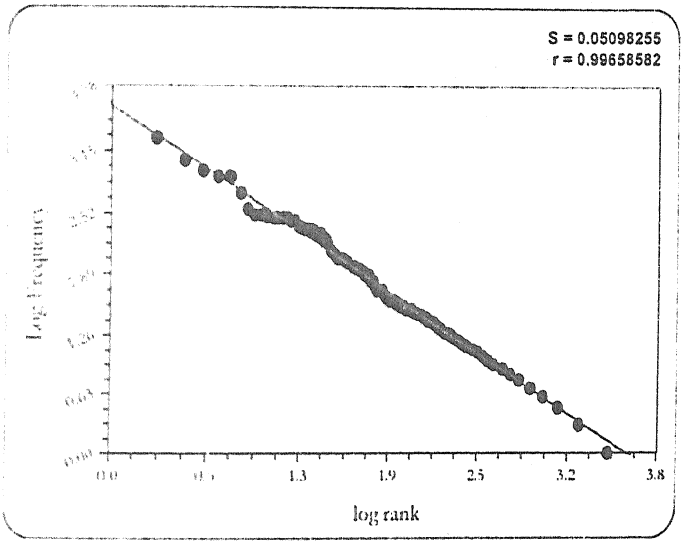


Figure 4.23: *Plot of log ranks vs. log frequency in "The Library"*

The Linear Fit y = a+bx for the log rank and log frequency data obtained coefficients as a = 3.60 and b = - 1.00. The residual[8] plot, which shows the difference between the data points and the model, evaluated at the data points is shown below.

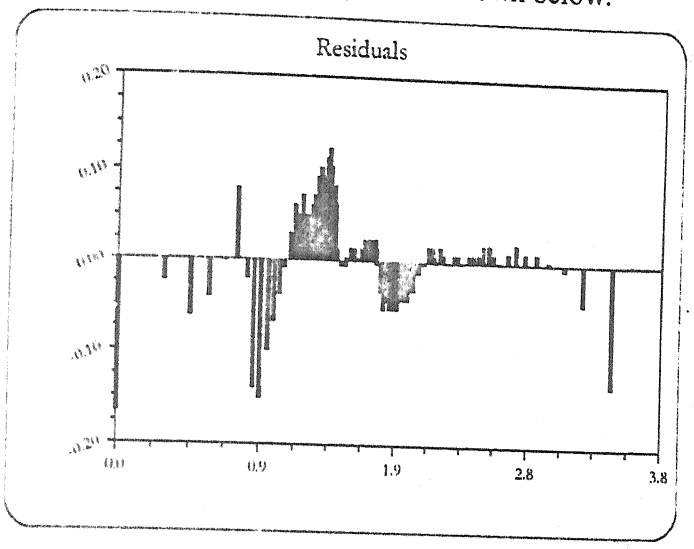


Figure 4.24: *Residual Plot for data points & model in "The Library"*

The result verified that Zipf's law is applicable in this text and for the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$), the coefficient c in this case is –1.00

[8] The residual at point k is defined by $\text{Residual}_k = y_k - f(x_k)$

Where y_k is the measured value at x_k, and $f(x_k)$ is the predicted value at x_k.

Section 7: Zipf's Law in Urdu Literature (*Bisat-e-Hyder*[38] *by Hyder Zaheer Ansari Hyder.*)

For this section, we have taken an e-text from the English version of the collection of Ghazal "Bisat-e-Hyder" by Hyder Zaheer Ansari Hyder.
(http://www.bisatehyder.indiaaccess.com/)
The choice of this text was done mainly due to the following reasons:

- Ghazals is a genre of music or poetry, which is essentially addressed to divine love. Two facets portray the ghazals: deep spirituality and passionate love. It is therefore very popular and representative of Urdu literature.
- English translation of this text was available so that analysis of the text was possible by Text Stat. This text is easily available on the web and is popular also. A testimony in this regard is a mail from the then president of USA, Mr. Bill Clinton, and "Thank you very much for your kind gift I appreciate your kind thoughtfulness and generosity. You have my best wishes".

The following statistics were obtained for the text under consideration.

Statistic for the document	
Number of words	4035
Number of sentences	529
Number of words per sentence	7.63
Number of syllables per word (approximate)	1.60
Flesch index[9]	63.63

The Flesch index of readability for this document is 63.63%. This means it is a fairly easy document to understand and is of the level of ninth standard. The prominent words, which were obvious to get more occurrences in this text were, "dil". "mujhko", "mohobbat", "baat", "dared", "yadh", "gum", "hyder", and "gazal" . Supporting words

[9] For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

like "hai", "main", "ki", "ka", "ke", "se", "yeh", "hain", "to", "ko", "ho" also obtained lot of occurrences.

The zipfian data for the text is obtained and is presented on the table given on the next page.

Rank	Freq	Log Rank	Log Freq
1	155	0.00	2.19
2	98	0.30	1.99
3	82	0.48	1.91
4	68	0.60	1.83
5	60	0.70	1.78
6	59	0.78	1.77
7	49	0.85	1.69
8	47	0.90	1.67
10	46	1.00	1.66
11	44	1.04	1.64
13	43	1.11	1.63
15	41	1.18	1.61
16	38	1.20	1.58
17	32	1.23	1.51
18	31	1.26	1.49
19	30	1.28	1.48
20	27	1.30	1.43
22	26	1.34	1.41
23	25	1.36	1.40
24	24	1.38	1.38
25	23	1.40	1.36
29	22	1.46	1.34
30	20	1.48	1.30
34	19	1.53	1.28
35	18	1.54	1.26
37	17	1.57	1.23
38	16	1.58	1.20
41	15	1.61	1.18
45	14	1.65	1.15
49	13	1.69	1.11
52	12	1.72	1.08
56	11	1.75	1.04
60	10	1.78	1.00
68	9	1.83	0.95
76	8	1.88	0.90
91	7	1.96	0.85
101	6	2.00	0.78
116	5	2.06	0.70
136	4	2.13	0.60
167	3	2.22	0.48
249	2	2.40	0.30
424	1	2.63	0.00

Table 4.19: *Zipfian data for the text (Bisat-e-Hyder by Hyder Zaheer Ansari Hyder)*

With this framework and preliminary analysis, we proceeded for the regression and curve fitting for this text.

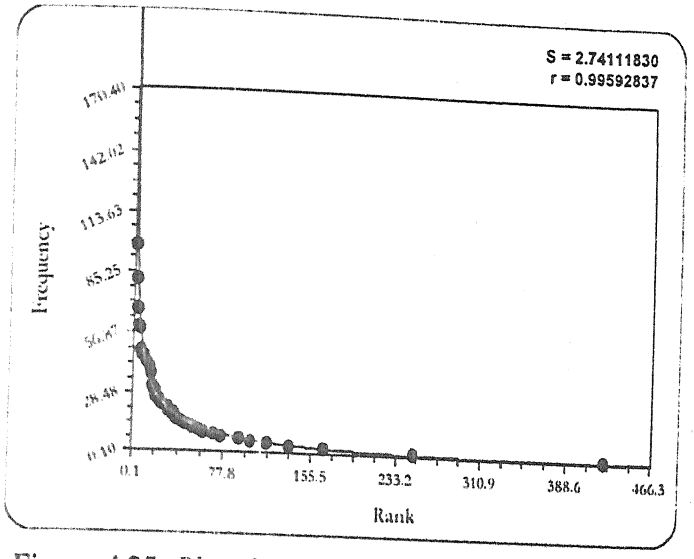


Figure 4.25: *Plot of ranks vs. Frequency in Bisat-e-Hyder*

Hoerl Model, $y = ab^x x^c$ with coefficient data, a = 151.06, b = 0.99 and c = -0.51 fitted this data in the manner which is shown in the graph given above.

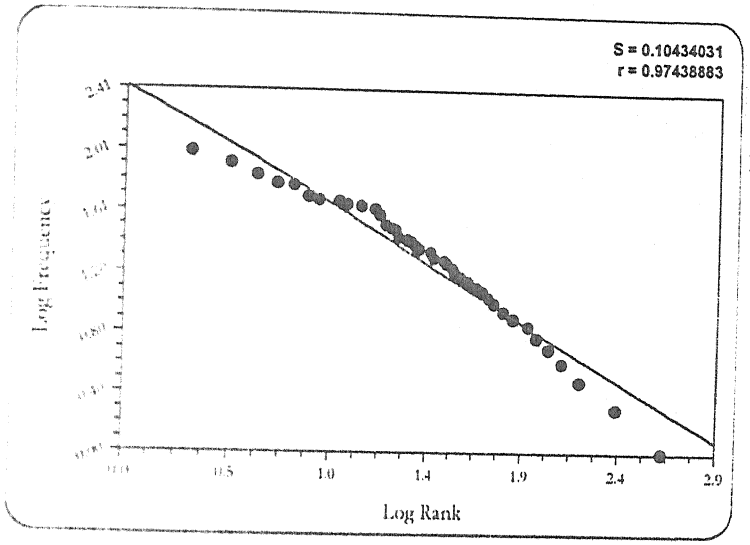


Figure 4.26: *Plot of log ranks vs. log frequency in Bisat-e-Hyder*

The Linear Fit y = a+bx for the log rank and log frequency data obtained coefficients as a = 2.43 and b = - 0.81. The residual[10] plot, which shows the difference between the data points and the model, evaluated at the data points is shown below.

[10] The residual at point k is defined by $Residual_k = y_k - f(x_k)$ Where y_k is the measured value at x_k, and $f(x_k)$ is the predicted value at x_k.

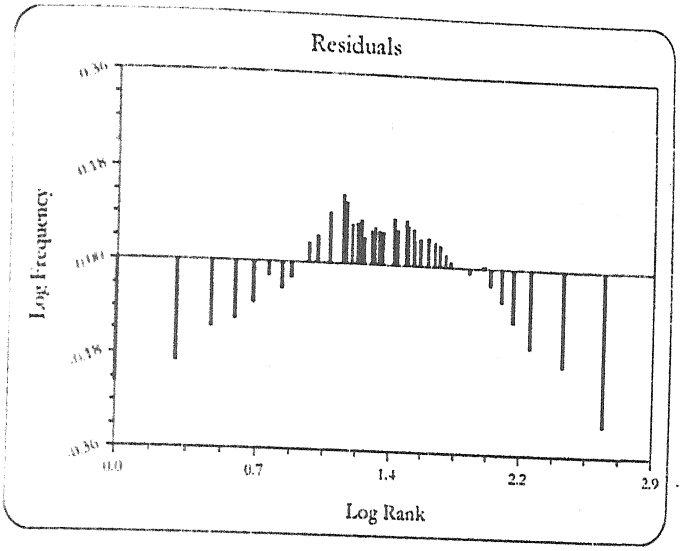


Figure 4.27: *Residual Plot for data points & model in Bisat-e-Hyder*

The result verified that Zipf's law is applicable in this text and for the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$) the coefficient c in this case is –0.81.

Section 8: Zipf's Law in Sanskrit Literature (*"Sri Vishnu Sahasranaamam[39]"*)

For this section we have taken, the Project Gutenberg E-Book of "Sri Vishnu Sahasranaamam", by Unknown. It is in Sanskrit and character set encoding is US-ASCII. This E-text was transcribed by N. Srinivasan and Karthik Krishnan and formatted by Maitri Venkat-Ramani. This e-text can be transliterated in Sanskrit using the ITRANS processing tool at http://sanskrit.gde.to/processing_tools/processing_tools.html.

The choice of this text was done mainly due to the following reasons:

- Sanskrit is one of the 22 official languages of India. According to Wikipedia, "Sanskrit is an Indo-European classical language of India and a liturgical language of Hinduism, Buddhism, and Jainism. It has a position in India and Southeast Asia similar to that of Latin and Greek in Europe, and is a central part of Hindu tradition".
- Sanskrit is mostly used as a ceremonial language in Hindu religious rituals in the forms mantras. The text taken here signifies this as it is addressed to Lord Vishnu.
- The following statistics were obtained for the text under consideration.

Statistic for the document	
Number of words	1411
Number of sentences	283
Number of words per sentence	4.99
Number of syllables per word (approximate)	3.33
Flesch index[11]	-79.97

The Flesch index of readability for this document is –79.97. This means it is a fairly difficult document to understand. We have found 1248 distinct word. This was expected also as the title under consideration is about synonyms of the names of lord Vishnu. Hence the repetition of words was not expected. The repetitions which are present are

[11] For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

basically the connecting words or explanatory words. The prominent words, which got more occurrences in this text, were as follows:

Word	Frequency
Ya	13
Cha	8
No	7
Sarva	7
Aum	4
Avyayah	4
Na	4
Naam	4
Paramam	4
Purushah	4
Vishnum	4
Vishnur	4
Yo	4

Table 4.20: *Prominent Sanskrit words & their frequency*

One can observe that supporting words like "ya", "cha", "no", "aum", "na", "yo" and "sarva" obtained lot of occurrences. The zipfian data for the text is obtained and is presented on the table given on the next page.

Rank	Frequency	Log Rank	Log Freq
1	13	0.00	1.11
2	8	0.30	0.90
3	7	0.48	0.85
5	4	0.70	0.60
14	3	1.15	0.48
23	2	1.36	0.30
110	1	2.04	0.00

Table 4.21: *Zipfian data for the text ("Sri Vishnu Sahasranaamam")*

With this framework and preliminary analysis, we proceeded for the regression and curve fitting for this text.

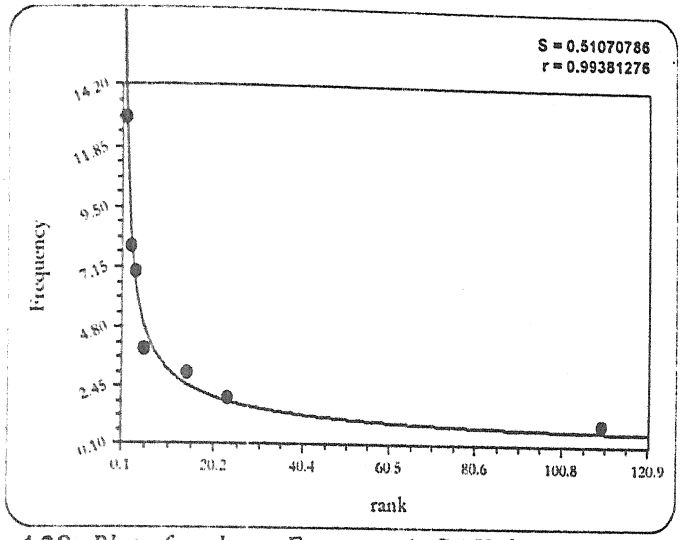


Table 4.28: *Plot of ranks vs. Frequency in Sri Vishnu Sahasranaamam*

Power Fit, $y = ax^b$ with coefficient data, a = 12.83 and b = -0.61 fitted this data in the manner which is shown in the graph given above.

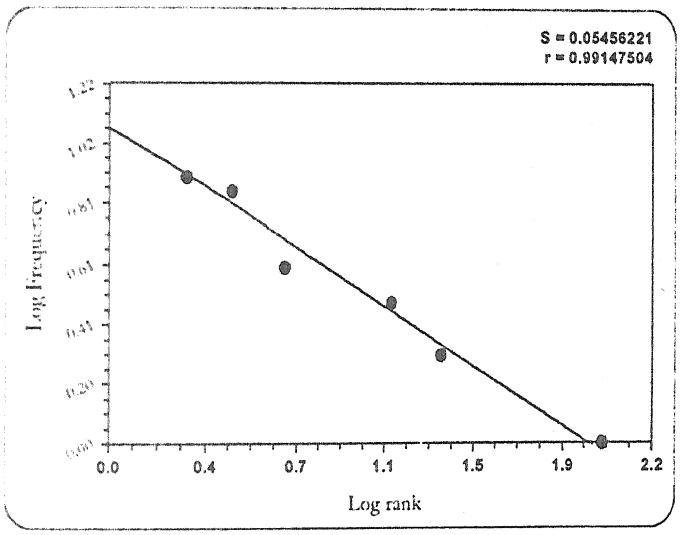


Figure 4.29: *Plot of log ranks vs. log frequency in Sri Vishnu Sahasranaamam*

The Linear Fit $y = a+bx$ for the log rank and log frequency data obtained coefficients as a = 1.07 and b = -0.54. The residual[12] plot, which shows the difference between the data points and the model, evaluated at the data points is shown below. Linear Fit: y=a+bx

[12] The residual at point k is defined by $Residual_k = y_k - f(x_k)$ Where y_k is the measured value at x_k, and $f(x_k)$ is the predicted value at x_k.

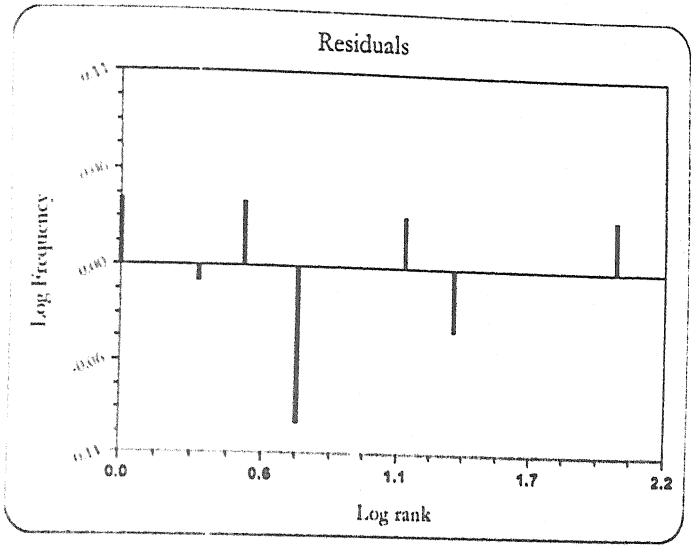


Figure 4.30: *Residual Plot for data points & model in Sri Vishnu Sahasranaamam*

The result verified that Zipf's law is not applicable in this text. For the Mandelbrot Zipf's law ($g(r) = a(r+b)^{c}$) the coefficient c in this case is –0.54 (which is not close to unity).

Section 9: Zipf's law and Flesch Readability Index

Zipf[8] (1949) in his work, "Human Behavior and the principle of least effort" viewed language as a "tool" that is shaped by its "jobs" in human society. Other works of Zipf were "Selective Studies and the Principle of Relative Frequency in Language" [16] which as published in 1932 and "Psycho-Biology of Languages" [17] which was published in 1935.

Many years after his death linguistics agreed that speakers simplify communication by using a small pool of words that they can retrieve quickly from their memory and listeners simplify communication by preferring words with a single and unambiguous meaning. This proved that Zipf's law is applicable in understanding human language.

Zipf searched for a principle of least effort that would explain the equilibrium between uniformity and diversity in usage of words. Most others searched for a probabilistic explanation. The burning question still remains- Do we have any new evidence that Zipf's explanation of principle of least effort is more correct than a statistical explanation?

Flesch Readability Index[30] on other hand has become a sort of a standard as far as the readability of the documents is concerned. At many places, it has become imperative to ascertain that the document/ forms have a fairly high value of Flesch Readability Index, so that it is understood by masses.

In this section, we have tried to investigate whether there is any relation between the Zipf's principle of least effort and the readability of the document.

Zipf's Law

Zipf formulated a law in 1930 that says frequency count (number of occurrence) of words in any text is inversely proportional to the rank of that word. In other words, the distribution of words adhered to a regular statistical pattern or "The probability of occurrence of words or other items starts high and tapers off exponentially. Thus, a few occur very often while many others occur rarely" (Black[18], 2000).

To further explain the basic form of the law,

frequency * rank has a inversely proportional relationship:

frequency * rank = constant or f * r = c

Zipf attributed this law as a consequence of "Principle of Least Effort". The Principle of Least Effort postulates that a person would like to communicate in such a way as to minimize his total effort. Altmann[19] (2002) commented that Zipf's ideas are the foundation stones of modern quantitative linguistics and his influence is not restricted to linguistics but incessantly penetrates other sciences. Mandelbrot[10] (1953) tried to discuss Zipf's law in terms of communication costs and explained that the communication costs increases as the number of words and their length grows. Ferrer-i-Cancho & Sole[20] (2001a) commented that many models of syntactic communication assume this law. It is an obvious ingredient for any theory of language evolution. According to Li[21] (2002), the number of times a word is used in written human languages and the frequency of usage are the variables that indulge in a Zipf's type distribution. Smith & Devine[22] (1985) found that legal texts also follows Zipf's law but in a little different manner. Francis & Kucera[23] (1964) applied the Zipf's law to the Brown corpus of 1 million words of American English. Le Quan Ha[24] et al. (2002) analyzed Zipf's law for large corpora in two languages, English (from the Wall Street journal) and Mandarin (from the People's Daily Newspaper and the Xinhua News Agency. Wang[25] (1989) presented Zipf's distribution of Chinese corpus and Wyllys[26] (1981) took a data set of 3907 English words. Sun[27] et al. (1999) commented, "Studies of word frequency have many interesting and potentially significant applications. For example this model could be used to evaluate a single article or an author's work. Assuming a reasonable level of skill among the writers whose works are the basis for our observations, we can use this model as a benchmark for assessing writer's language skills". Gelbukh & Sidorov[28] (2001) observed that the coefficients of Zipf law are different for different languages. Ferrer-I-Cancho & Sole[29] (2001b) showed that the co-occurrence of words in sentences relies on the network structure of the lexicon. They analyzed the properties in depth and commented that human language can be described in terms of a graph of word interactions.

Flesch Readability Index[30]

For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ \ Index\ = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

According to Wikipedia, the free encyclopedia, a syllable is a unit of organization for a sequence of speech sounds. Syllables are often considered the phonological "building blocks" of words. They can influence the rhythm of a language.

Flesch readability index can be related to the educational level of the audience. For example a score of 91-100 can be easily comprehensible by a 5th grade student, a score of 51-60 understandable by a High School student, a college graduate will be able to comprehend a document with score 31-50 and a document with score less than 0 can be understood by a Law School Graduate only.

Research Question

If a document has a high Flesch Readability Index, then whether the Zipf's curve will fit this document in a better manner. In other words, if a document is fairly easy to understand, then whether it will follow the Zipfian distribution? Whether Zipf's law is applicable in understanding the human language? Can it be used as benchmark for assessing a writer's skill?

Findings

Appendix I illustrates the documents with related statistics on number of words in the document, Flesch Readability Index, Zipf's coefficient, number of sentences, number of syllables per word and the number of words per sentence. On the basis of this, we tried to group documents primarily with respect to values of the Flesch Readability Index and the Zipf's coefficient.

Type I documents were those with a very high negative Flesch Readability Index and also a poor Zipf's coefficient. This was understandable as first one was a group of German words taken from English-German Business dictionary. The second one was "slokas" from Sanskrit language that follows a very different style. These documents thus possessed very less words per sentence and more syllables per sentence, resulting in highly negative values of Flesch Readability Index.

Type II document was English words taken from English-German Business dictionary. We expected that with improvement in Flesch Readability Index, the performance of Zipf's coefficient will also improve, but that did not happen. Type III document was a classic story from Hindi literature. It although have a low Flesch Readability Index but had an improved Zipf's coefficient as compared to type I and II documents.

Type IV documents had excellent value of Zipf's coefficient and also good readability (understandable by a high school level reader). This tend to show that Zipf's law is applicable in documents that have on an average 1.5 syllables per word and have 5-8 words per sentence.

Type V documents however nullified the claim that was found in type IV documents. Almost all these documents have Zipf's coefficient ranging from -1.20 to -1.37, but had variable readability indexes ranging from 46-80. No trend has either been found in the syllables per word and words per sentence.

Figure 1 shows the relation between Zipf's coefficients and Flesch Readability Index. One can easily visualize the type of documents here. Can we conclude that if a document has a highly negative readability index it is bound to have a bad Zipfian fit? The curve that fitted this type of distribution is Sinusoidal Fit: y=a+b*cos(cx+d) with coefficient data a =-0.83, b =0.40, c =0.03 and d =1.31 with standard error = 0.16 and correlation coefficient = 0.79.

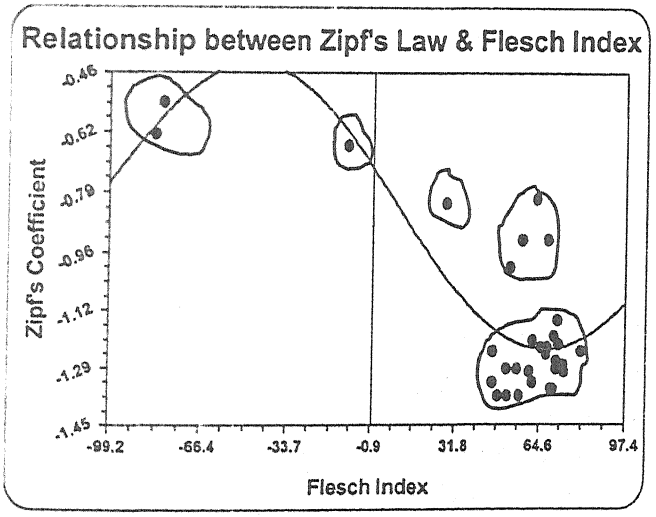


Figure 4.31: *Showing relation between Zipf's coefficients and Flesch Readability Index.*

Conclusion

Although we have tried to make the sample as diverse as possible, we finally found that 21 documents are belonging to type V. These documents have different readability indices, belong to a different genre and belong to different time periods but have almost similar value for the Zipf's coefficient. This indicates that readability has little to do with the Zipf's coefficients.

This led us to go back to our research question that if a document has a high Flesch Readability Index, then whether the Zipf's curve will fit this document in a better manner. Type IV documents partially demonstrate this as they have excellent value of Zipf's coefficient and also good readability. Type V documents also have Zipf's coefficients that are not too bad but these coefficients are constant while readability varies from 46 to 80. Type I documents however proved that a poor Zipf's coefficient indicates high negative Flesch Readability Index.

Coming to the next research question, whether Zipf's law is applicable in understanding the human language? Can it be used as benchmark for assessing a writer's skill? The findings in this communication refute this claim. It is because of the fact that Zipf's principle of least effort says that a writer simplifies communication by using a small pool of words from their memory. This would mean that these communications ought to have good readability indices too. So, all those documents that have good readability coefficients should have good Zipf's coefficient also. This however is not reflected in the findings. Hence the contradiction that Zipf's law is applicable in understanding human language. This is also contradictory to Sun[27] et al. (1999) comment that "we can use this model as a benchmark for assessing writer's language skills".

In conclusion, we can say that probably more data-sets need to be taken to formalize these findings.

Table 4.22: *Document Statistics and Zipf's Law*

S.no	File Name	No- of Words	Flesch Index	Zipf's Coefficient	No- of Sentences	Syllables per word	Word per sentence	Type
1	Eng-ger-busDictionary.txt	9107	-82.83	-0.63	5792	3.41	1.57	1
2	sanskritwork.txt	1411	-79.97	-0.54	283	3.33	4.99	1
3	Eng-ger-busDictionary.txt	10089	-9.95	-0.66	5763	2.54	1.75	2
4	eidgaah.txt	4951	28.74	-0.82	505	1.99	9.8	3
5	library s.txt	37498	53.33	-1	5037	1.73	7.44	4
6	aladdin ger.txt	17686	57.9	-0.92	3536	1.7	5	4
7	aladdin eng.txt	5319	68.23	-0.92	661	1.54	8.05	4
8	urdu.txt	4035	63.63	-0.81	529	1.6	7.63	4
9	jefferson-autobiography-73.txt	40648	48.76	-1.37	5326	1.78	7.63	5
10	wollstonecraft-maria-196.txt	45874	52.75	-1.37	5426	1.72	8.45	5
11	franklin-autobiography-244.txt	68157	57.27	-1.37	7270	1.66	9.38	5
12	chaucer-canterbury-102.txt	99403	69.18	-1.35	13578	1.54	7.32	5
13	augustine-confessions-276.txt	176014	69.99	-1.35	22974	1.53	7.66	5
14	mill-subjection-217.txt	45240	46.75	-1.33	5108	1.79	8.86	5
15	Arabian nights entertainments.txt	90768	62.41	-1.33	10672	1.61	8.51	5
16	aristotle-meteorology-80.txt	43470	60.99	-1.3	5030	1.62	8.64	5
17	freud-young-763.txt	72133	74.4	-1.3	11160	1.49	6.46	5
18	berkeley-treatise-177.txt	36342	52.17	-1.29	4115	1.72	8.83	5
19	locke-concerning-111.txt	53786	56.56	-1.29	5732	1.66	9.38	5
20	barrie-peter-277.txt	47885	71.56	-1.29	6906	1.52	6.93	5
21	bunyan-pilgrims-304.txt	57122	73.73	-1.28	7241	1.48	7.89	5
22	anonymous-beowulf-543.txt	27129	71.35	-1.27	4173	1.52	6.5	5
23	dickens-christmas-125.txt	21818	67.75	-1.25	3301	1.56	6.61	5
24	hiroshima nagasaki.txt	25341	46.75	-1.24	3313	1.8	7.65	5
25	twain-tom-40.txt	24486	80.99	-1.24	3564	1.41	6.87	5
26	lucretius-on-395.txt	75386	65.78	-1.23	10549	1.58	7.15	5
27	keats-endymion-484.txt	31962	68.19	-1.23	4847	1.56	6.59	5
28	365 foriegn dishes.txt	27891	72.03	-1.22	4424	1.52	6.3	5
29	The arctic queen.txt	16703	62.09	-1.21	2451	1.63	6.81	5
30	shakespeare-hamlet-25.txt	33098	70.42	-1.2	4931	1.53	6.71	5
31	shakespeare-romeo-48.txt	26784	71.97	-1.15	3854	1.51	6.95	5

Section 10: Zipf's Law and Principle of Least effort

Zipf attributed his law as a consequence of "Principle of Least Effort". The Principle of Least Effort postulates that a person would like to communicate in such a way as to minimize his total effort. In other words, a person will tend to "minimize" the probable average of his work-expenditure (over time), meaning use of least amount of work. Principle of Least Effort is relevant even today. However, it was criticized by Rapaport[31] (1957) on the basis that although Zipf's arguments are plausible in a great variety of situations, they are not suitable for generalizations.

Zipf[8] (1949) in his work, "Human Behavior and the principle of least effort" viewed language as a "tool" that is shaped by its "jobs" in human society. The purpose of this book, which was an introduction to human ecology, "is to establish the Principle of least effort as the primary principle that governs our entire individual and collective behavior of all sorts".

Chai Kim[32] (1982) investigated the extent to which the principle of least effort as advanced by Zipf provided a theoretical basis for identifying and updating descriptors of science/technology and social sciences. He found that "the relative frequency of occurrence of the descriptors of social sciences conformed to the theoretical distribution of Zipf while that of science/technology did not".

In this section we will try to view the above facts on the basis of the 31 documents that we have analyzed for checking the robustness of Zipf's law. We have thus collected data on the following parameters:

- *No- of Word*: This is the total number of words in a document
- *Sum of frequency*: Sum of frequencies is the sum of rank frequencies of the Zipfian data of the document. When shown as a percentage of the total words it reveals the *contain%* of the document.
- *Unique words*: This is the count of unique words in a document.
- *Least effort %*: This is the ratio (percentage) of the unique words to the total number of words in the data. It reveals the "effort" that the writer has done in

communicating his ideas. The smaller the percentage the less is the effort of the writer.

- *Contain%:* This is the ratio (percentage) of the sum of rank frequencies of the Zipfian data of the document to the total number of words in the data. It reveals the amount by which the Zipfian data is able to capture the document. The higher the percentage the more is the better the containment of document in the Zipfian data.

- *Flesch Index*: For a given document, the Flesch readability index is an integer indicating how difficult the document is to understand, with lower numbers indicating greater difficulty.

$$Flesch\ Index = 206.835 - 84.6 * \frac{syllables}{words} - 1.015 * \frac{words}{sentences}$$

- *Zipf's Coefficient*: It refers to Mandelbrot generalization of Zipf's law that the slope is -1. Mandelbrot (1952, 1964) assumed that the aim of language is to transmit the most information per symbol with the least effort. It is expressed by the relationship $f(r) = k(r+c)^{-\theta}$ where, *f(r)* is the rank frequency and r is the rank of the word. 'c' and 'θ' are constants, 'c' improves the fit for small r and the exponent 'θ' improves the fit for large r. Here the Zipf's coefficient refers to the exponent 'θ'.

Now with this data in hand, we tried to apply factor analysis. Factor analysis attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables. Factor analysis is often used in data reduction, by identifying a small number of factors, which explain most of the variance observed in a much larger number of manifest variables.

File Name	No- of Words	Sum (freq)	Unique words	Least effort %	Contain%	Flesch Index	Zipf's Coeff.
augustine-confessions-276.txt	176014	117820	9629	5.47	66.94	69.99	-1.35
365 foriegn dishes.txt	27891	16586	1581	5.67	59.47	72.03	-1.22
freud-young-763.txt	72133	47288	4486	6.22	65.56	74.4	-1.3
Arabian nights entertainments.txt	90768	55204	6464	7.12	60.82	62.41	-1.33
aristotle-meteorology-80.txt	43470	27171	3186	7.33	62.51	60.99	-1.3
bunyan-pilgrims-304.txt	57122	36420	4274	7.48	63.76	73.73	-1.28
librarys.txt	37498	18037	2809	7.49	48.10	53.33	-1
locke-concerning-111.txt	53786	34675	4169	7.75	64.47	56.56	-1.29
sanskritwork.txt	1411	38	1248	88.45	2.69	-79.97	-0.54
berkeley-treatise-177.txt	36342	21858	3222	8.87	60.15	52.17	-1.29
chaucer-canterbury-102.txt	99403	59892	8854	8.91	60.25	69.18	-1.35
aladdin ger.txt	17686	7234	1633	9.23	40.90	57.9	-0.92
franklin-autobiography-244.txt	68157	39154	6496	9.53	57.45	57.27	-1.37
aladdin eng.txt	5319	2521	524	9.85	47.40	68.23	-0.92
lucretius-on-395.txt	75386	41078	7446	9.88	54.49	65.78	-1.23
barrie-peter-277.txt	47885	28498	4788	10.00	59.51	71.56	-1.29
twain-tom-40.txt	24486	14493	2455	10.03	59.19	80.99	-1.24
urdu.txt	4035	1280	424	10.51	31.72	63.63	-0.81
mill-subjection-217.txt	45240	26259	4885	10.80	58.04	46.75	-1.33
wollstonecraft-maria-196.txt	45874	25319	5940	12.95	55.19	52.75	-1.37
shakespeare-romeo-48.txt	26784	14311	3541	13.22	53.43	71.97	-1.15
jefferson-autobiography-73.txt	40648	21556	5497	13.52	53.03	48.76	-1.37
hiroshima nagasaki.txt	25341	12168	3448	13.61	48.02	46.75	-1.24
shakespeare-hamlet-25.txt	33098	18264	4542	13.72	55.18	70.42	-1.2
anonymous-beowulf-543.txt	27129	12315	3744	13.80	45.39	71.35	-1.27
Eng-ger-busDictionary.txt	10089	1690	1402	13.90	16.75	-9.95	-0.66
Eng-ger-busDictionary.txt	9107	645	1378	15.13	7.08	-82.83	-0.63
dickens-christmas-125.txt	21818	10723	3695	16.94	49.15	67.75	-1.25
keats-endymion-484.txt	31962	14182	5521	17.27	44.37	68.19	-1.23
The arctic queen.txt	16703	6776	3482	20.85	40.57	62.09	-1.21
eidgaah.txt	4951	1698	1497	30.24	34.30	28.74	-0.82

Table 4.23: *Least effort percentage and contain percentage of the documents*

The following descriptive statistics was obtained when we proceeded for the factor analysis with the four variables; least effort % (LE_PER); contain%(DEF_PER); Flesch Index (F_INDEX) & Zipf's Coefficient (ZIPF_C).

	Mean	Std. Deviation
DEF_PER	49.2219	16.2034
F_INDEX	50.7394	39.0706
LE_PER	14.0561	14.6947
ZIPF_C	-1.1535	.2388

Table 4.24: *Descriptive Statistics of Variables*

The Correlation Matrix is obtained suggests that there is strong negative correlation between Contain % and the Zipf's coefficient, on the other hand there is a strong correlation between Contain% and the Flesch readability index. There is a weak correlation between least effort % and the Zipf's law.

	DEF_PER	F_INDEX	LE_PER	ZIPF_C
DEF_PER	1.000	.849	-.661	-.907
F_INDEX	.849	1.000	-.667	-.752
LE_PER	-.661	-.667	1.000	.556
ZIPF_C	-.907	-.752	.556	1.000

Table 4.25: *Correlation matrix of variables*

Bartlett's test of sphericity tests whether the correlation matrix is an identity matrix, which would indicate that the factor model is inappropriate. So the null hypothesis that the correlation matrix is an identity matrix is tested and the following results were obtained (χ^2= 102.21, df = 6, p<0.00). The p-values found here is significant hence we reject the null hypothesis that the inter-correlation matrix comes from a population in which the variables are non-collinear (i.e. an identity matrix). The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy tests whether the partial correlations among variables are small.

The KMO statistic is 0.769 that means we can conclude that the degree of common variance among the variables is all right and the factors extracted will account for fare amount of variance.

We have also calculated communalities, which is the proportion of the total variance of a variable accounted for by the common factors in a factor analysis. All the variables have

communality above 0.95 with LE_PER having the highest communality (1.00) and DEF_PER having the lowest (0.955). The extraction method was based on Principal Component Analysis.

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.209	80.232	80.232	3.209	80.232	80.232
2	.491	12.274	92.506	.491	12.274	92.506
3	.229	5.737	98.243	.229	5.737	98.243
4	0.007028	1.757	100.000			

Extraction Method: Principal Component Analysis.

Table 4.26: *Components and % of variance they explain in Factor Analysis*

The following factors were obtained:

Component Matrix			
	Component		
	1	2	3
DEF_PER	.961←	.171	-5.171E-02
F_INDEX	.916←	1.269E-02	.396
ZIPF_C	-.905	-.326	.230←
LE_PER	-.792	.596←	.132

Extraction Method: Principal Component Analysis.

Table 4.27: *Factors obtained by Principal Component Analysis*

Factor1 (DEF_PER & F_INDEX) has 2 variables. Factor 2 (LE_PER) has one variable and factor 3 (ZIPF_C) has 1 variable. The first two factors are explaining almost 92% of the total variance. The following Scree plot has been obtained:

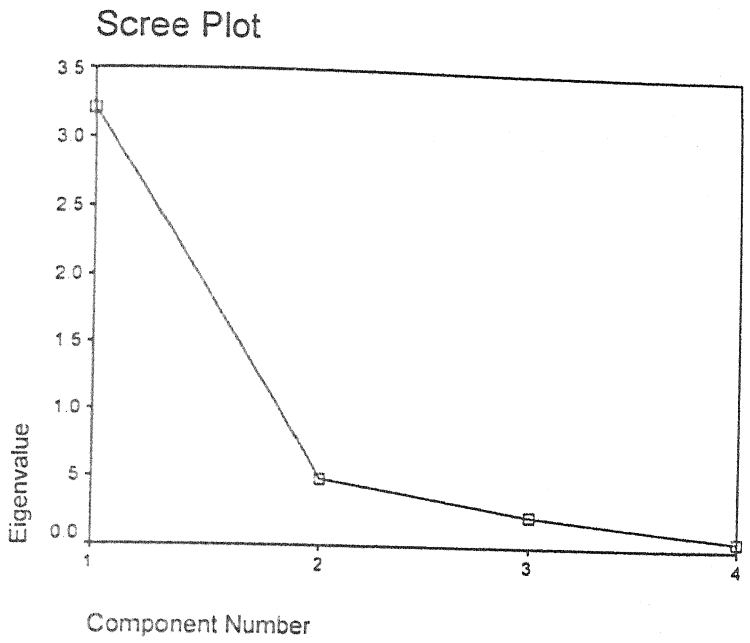


Figure 4.32: *SCREE Plot of factors*

Now we are able to comment that contain% and the Flesch readability index together with the least effort% are explaining 92% of the variance of the data. The Zipf's coefficient is a redundant data here. To probe more into this we have taken these three as predictor variables in a multiple regression analysis to predict the dependent variable Zipf's coefficient.

Model Summary

	R	R-Sq	Adj.R-Sq	S.E. Estimate	Change Statistics				
Model					R-Sq Change	F Change	df1	df2	Sig. F Change
1	.907	.823	.816	.1023	.823	134.449	1	29	.000

a Predictors: (Constant), DEF_PER
b Dependent Variable: ZIPF_C

Table 4.28: *Model Summary in Multiple Regression Analysis*

ANOVA

Model		SS	df	Mean Square	F	Sig.
1	Regression	1.407	1	1.407	134.449	.000
	Residual	.303	29	1.046E-02		
	Total	1.710	30			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.496	.060		-8.31	.000		
	DEF_PER	-1.336E-02	.001	-.907	-11.59	.000	1.000	1.000

Excluded Variables

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	LE_PER	-.078	-.744	.463	-.139	.563	1.778	.563
	F_INDEX	.063	.422	.676	.080	.280	3.572	.280

a Predictors in the Model: (Constant), DEF_PER

b Dependent Variable: ZIPF_C

Table 4.29: *Multiple Regression Analysis (ANOVA Table)*

The model found in the analysis suggests that DEF_PER alone explains 82.3% of Zipf's coefficient. Adjusted R^2 for the model is 0.816. R^2 is adjusted to reflect the model's goodness of fit for the population. The net effect of this adjustment is to reduce R^2 from 0.823 to 0.816, thereby making it comparable to other R^2's in case other models are also found.

Standard error of the model is 0.1023. This is the standard deviation of actual values of Y about the estimated Y values. Analysis of Variance measures whether or not the equation represents a set of regression coefficients that, in total, are statistically significant from zero. The critical value for F is found to be 134.44, which is significant at less than 0.05 level of significance at1, 29 df. Regression Coefficients for the model and the

unstandardized regression coefficients for the equation are given in the table. The equation may be constructed as

$$ZIPF\_C = -.496 - 0.01336\ DEF\_PER$$

Or

$$ZIPF\_C = -0.907\ DEF\_PER$$

Since variability inflation factor (VIF) is 1.0, multicollinearity is not a problem. The equation above suggests that the Zipf's coefficient of the documents can be predicted well by the variable DEF_PER or the contain% only and the variable LE_PER or least effort is not required.

The result thus came close to the finding of Rapaport[31] (1957) that although Zipf's arguments are plausible in a great variety of situations, they are not suitable for generalizations. It also supported Chai Kim[32] (1982) that "the relative frequency of occurrence of the descriptors of social sciences conformed to the theoretical distribution of Zipf while that of science/technology did not". In this data, documents like English German business dictionary, Sanskrit text and Urdu text behave differently than the Zipfian distribution.

References

1. Martindale, Colin. & Konopka, Andrzej. K. (1996). Oligonucleotide frequencies in DNA follow a Yule distribution. *Computer & Chemistry,* 20(1): 35-38.
2. Perline, Richard. (1996). Zipf's law, the central limit theorem, and the random division of the unit interval. *Physical Review E,* 54(1): 220-223.
3. Laherrere, Jean. & Sornette, D. (1998). Stretched exponential distributions in nature and economy: 'Fat tails' with characteristic scales. *European Physical Journal* B2: 525-539.
4. Miller,G. A. & Newman, E. B. (1958). Tests of a statistical explanation of the rank-frequency relation for words in written English. *American Journal of Psychology*, 71, 209-218
5. Rousseau, R. & Zhang, Qiaoqiao. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics*, 24(2): 201-220.
6. Landini, G. (1997). Zipf's laws in the Voynich manuscript. http://sun1.bham.ac.uk/G.Landini/evmt/zipf.htm
7. Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory,* 38(6): 1842-1845.
8. Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley, 1949.
9. Chen, Ye-Sho. & Leimkuhler, Ferdinand. F. (1987). Analysis of Zipf's Law: An index approach. *Information Processing and Management,* 23(3): 171-182.
10. Mandelbrot, B. (1953). An information theory of the statistical structure of language. *Proc. Symposium on Applications of Communication Theory,* September 1952; London: Butterworth; 486-500
11. Milenkovic, Milan. (1997). *Operating System- Concept & Design.* Second edition, Tata McGraw Hill, New Delhi.
12. Rousseau, Ronald. (2001). Evolution in time of the number of hits in keyword searches on the Internet during one year, with special attention to the use of word Euro", *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, Sydney, Vol2, pp 619-627.
13. Lee Breslau., Pei, Cao., Li, Fan., Graham, Phillips. & Scott, Shenkar. (1999). Web Caching and Zipf-Like Distributions: Evidence and Implications, *IEEE INFOCOM*, Vol. XX. NO- Y.
14. Chao, Dennis. & D'haeseleer, Patrik. (2001). The distribution of Variable-Length Phatic Interjectives on the World Wide Web, *UNM Computer Science Department Tech Report* TR-CS-2001-23.
15. Bar-Ilan, J. (2001). Data Collection Methods on the Web for Informetric purposes- A Review & Analysis. *Scientometrics*, 50, 7-32

16. Zipf G.K. (1932), *Selective Studies and the Principle of Relative Frequency in Language*.

17. Zipf G.K. (1935). *Psychobiology of Languages*, Houghton-Mifflin, 1935; MIT Press.

18. Black, Paul. E. (2000). Zipf's Law: Definition, at http://hissa.nist.gov/dads/HTML/zipfslaw.html site accessed on 1/29/01.

19. Altmann, Gabriel. (2002). Zipfian linguistics . *Glottometrics* 3, pp 19-26, 2002

20. Ferrer-i-Cancho, Ramon. & Sole, Richard. V. (2001a). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics* 8:165-174.

21. Li, Wentian. (2002). Zipf's Law Everywhere. *Glottometrics*, 5, 2002, 14-21

22. Smith, F. J. & Devine, K. (1985). Storing and Retrieving Word Phrases. *Information Processing & Management*, Vol. 21, No. 3, pp 215-224.

23. Francis, W. N. & Kucera, H. (1964). *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University, Providence, Rhode Island

24. Le Quan, Ha , Sicilia-Garcia, E. I., Ming, J. & Smith, F. J. (2002). Extension of Zipf's Law to Words and Phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, pages 315-320, Taipei, Taiwan.

25. Wang, C. (1989). Zipf's distribution of Chinese corpus, *Information Sciences*, 10, 1-8.

26. Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law, *Library Trends*, 30, 53-64

27. Sun, Qinglam., Shaw, D. & Davis, C. H. (1999). A model for estimating the occurrence of same frequency word and the boundary between the high and low frequency words in texts. *Journal of the American Society for Information Science*, Mar 1999; 50, 3

28. Gelbukh, Alexander. & Sidorov, Grigori. (2001). Zipf and Heaps Law's Coefficients Depend on Language. *Proc. CICLing-2001*, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335.

29. Ferrer-i-Cancho, Ramon. & Sole, Ricard. V. (2001b). The small world of human languag", *Proc. R. Soc. Lond.* B (2001) 268, 2261-2265.

30. http://users.dickinson.edu/~braught/courses/cs132f02/labs/lab07.html for information about Flesh Index. Site accessed on 02/06/2006.

31. Rapaport, Anatole. (1957). The Stochastic and the 'Teleological' Rationales of Certain Distributions and the So-called Principle of Least Effort, *Behav. Sci.*, 2, 150.

32. Chai, Kim. (1982). Retrieval Language of Social Sciences and Natural Sciences: A statistical Investigation. *Journal of the American Society for Information Sciences*, Jan 1982; 33, 1; ABI/Inform Global, Page 3.

33. The Project Gutenberg e-text of "Aladdin and the Wonder Lamp" (http://www.gutenberg.org/)

34. The Project Gutenberg e-text of "Aladdin und die Wunderlampe", by Ludwig Fulda (http://www.gutenberg.org/)

35. The Project Gutenberg E-text of "Mr. Honey's Small Business Dictionary (English-German)" by Winfred Honig.

36. The IIT Kanpur's e-text of roman version of "Eidgaah" by Munshi Prem Chand (http://www.munsipremchand.iitk.ac.in/auther.html)

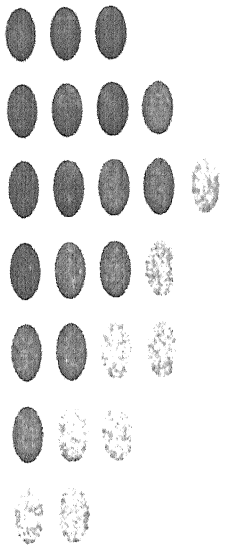
37. The Project Gutenberg e-text of "The Library", by Andrew Lang #20 (http://www.gutenberg.org/)

38. The e-text from the collection of Ghazal "Bisat-e-Hyder" by Hyder Zaheer Ansari Hyder. (http://www.bisatehyder.indiaaccess.com/)

39. The Project Gutenberg e-Book of "Sri Vishnu Sahasranaamam" (http://www.gutenberg.org/)

Chapter 5

Discussion



Discussion

According to George Bernard Shaw[1], Irish literary Critic and Essayist who won the Nobel Prize for Literature in 1925,

> *"New opinions often appear first as jokes and fancies, then as blasphemies and treason, then as questions open to discussion and finally as established truths".*

When we started working on this problem, the work looked like a mammoth task. There are several reasons for this. First and foremost was the popularity of Zipf's law and its applicability in a plethora of activities. Secondly, there were a gigantic number of exceptionally eminent people who have worked on works of Zipf.

We started by setting very modest goals for us these objectives are as follows:

- To find the interrelationships between the rank and the frequency of a word in selected literatures.
- To test whether the Zipf's law can be applied in these literatures.
- To do an inter-literature comparison of the applicability of Zipf's law.
- To do mathematical modelling & validation of the model through the collected data.

And we set the following hypothesis

1. The principle of least effort is a universal phenomenon.
2. All writers would follow an economy in the use of words irrespective of the languages concerned.
3. The rank-frequency distribution of words would be similar in all languages.

Let us discuss the analysis of the documents that we have collected for this work. For this task we have divided the discussion in three sections that addresses to the objectives mentioned above.

Interrelationships between the rank and the frequency

This section is devoted to the first objective that was to find the interrelationships between the rank and the frequency of a word in selected literatures. To probe into this, the documents were subjected to curve fitting. The Zipf's coefficient was obtained by fitting Zipf-Mandelbrot equation or the shifted power curve. Simultaneously, rank and frequency data was also subjected to curve fitting and we obtained the following major categories of distributions that were able to describe the rank-frequency distribution of the documents.

- Yield density family – Bleasdale and Harris Models
- Power Law family – Hoerl, Modified Hoerl, Power models
- Exponential family – Vapour Pressure models
- Sigmoidal family – MMF & Weibull models

As we can see four model families were used and all were able to explain around 97% - 99% of the variance. The big question which automatically arises is: - How to compare the appropriateness for the different type of functions fitted to this data. Should one just fit all the commonly used functions and see which one fits the data "the best". A good analysis requires robust techniques in assessing and empirically developing the model. The data is never wrong and thus "Statistics" should "speak" for the data. One should not lead by assumption but should try empirical evidence. The data itself suggests as to how & in what form the model is to be used. In summary, we have not adhered to pre-specifying the model but tried to develop the model by keeping it simple and parsimonious. Nobody can claim that a particular model is the true equation of the data in question as the true equation is only known to GOD and hence it is said that "All models are wrong, some are useful!"

Following is the classification of documents according to the distributions. These models gave good parameter values and fit statistics, but nobody was able to put forward a generalized rule or procedure.

The yield-density models are widely used to model the relationship between the yield of a crop and the spacing or density or planting. If the response is such that as density (x)

increases, but the yield (y) approaches a fixed value, the relationship is asymptotic. If the response is such that there is a distinct optimum as the density increases, the relationship is parabolic.

The documents like Aladdin (English), Aladdin (German), and English-German-Business Dictionary (English & German words) were the four documents that showed behavior like this family. Following are the fit parameters for this type.

Bleasdale Model, $y = (a + bx)^{\frac{-1}{c}}$

File Name	No- of Words	Unique words	Zipf's Coeff.	Relationship between Rank & Frequency	a	b	c
aladdin eng.txt	5319	524	-0.92	Bleasdale Model	0.15	0.01	0.36
aladdin ger.txt	17686	1633	-0.92	Bleasdale Model	0.06	0.00	0.47

Table 5.1: *Documents where the rank & frequency relationship followed Bleasedale model*

Harris Model, $y = \frac{1}{a + bx^c}$

File Name	No- of Words	Unique words	Zipf's Coeff.	Relationship between Rank & Frequency	a	b	c
Eng-ger-busDictionary.txt	10089.00	1402.00	-0.66	Harris Model	0.12	0.12	0.09
Eng-ger-busDictionary.txt	9107.00	1378.00	-0.63	Harris Model	-0.01	0.02	0.43

Table 5.2: *Documents where the rank & frequency relationship followed Harris model*

The second class of distribution was Sigmoidal Family. These "S-shaped" growth curves are common in a wide variety of applications such as biology, engineering, agriculture, and economics. These curves start at a fixed point and increase their growth rate monotonically to reach an inflection point. After this, the growth rate approaches a final value asymptotically. This family is actually a subset of the Growth Family, but is separated because of their distinctive behavior. Many documents were found to adhere to

this family of distributions. As many as 16 documents (out of 31) wee found to come from this family. Fit parameters for these documents are given below:

Weibull Model, $y = a - be^{-cx^d}$

File Name	No- of Words	Unique words	Zipf's Coeff.	Relationship between Rank & Frequency	a	b	c	d
jefferson-autobiography-73.txt	40648	5497	-1.37	Weibull Model	4519.53	4534.42	1.23	-1.04
wollstonecraft-maria-196.txt	45874	5940	-1.37	Weibull Model	2551.46	2592.51	3.08	-1.02
augustine-confessions-276.txt	176014	9629	-1.35	Weibull Model	10151.81	10324.38	1.76	-0.81
berkeley-treatise-177.txt	36342	3222	-1.29	Weibull Model	2111.22	2180.72	2.16	-0.83
locke-concerning-111.txt	53786	4169	-1.29	Weibull Model	3967.59	4007.99	1.80	-0.92
bunyan-pilgrims-304.txt	57122	4274	-1.28	Weibull Model	2487.01	3081.75	2.13	-0.80
anonymous-beowulf-543.txt	27129	3744	-1.27	Weibull Model	7606.39	7633.55	0.28	-0.84
keats-endymion-484.txt	31962	5521	-1.23	Weibull Model	1318.18	1341.75	2.81	-0.95
The arctic queen.txt	16703	3482	-1.21	Weibull Model	3311.38	3345.22	0.32	-0.72
shakespeare-romeo-48.txt	26784	3541	-1.15	Weibull Model	830.34	940.91	2.77	-0.63
freud-young-763.txt	72133	4486	-1.30	Weibull Model	2381.06	2467.13	5.22	-0.92

Table 5.3: *Documents where the rank & frequency relationship followed Weibull model*

MMF Model, $y = \frac{ab + cx^d}{b + x^d}$

File Name	No- of Words	Unique words	Zipf's Coeff.	Relationship between Rank & Frequency	a	b	c	d
franklin-autobiography-244.txt	68157	6496	-1.37	MMF Model	4301.03	3.66	-43.47	1.10
chaucer-canterbury-102.txt	99403	8854	-1.35	MMF Model	5528.24	3.08	-106.65	0.95
Arabian nights entertainments.txt	90768	6464	-1.33	MMF Model	77313.18	0.09	-122.78	0.74
dickens-christmas-125.txt	21818.00	3695.00	-1.25	MMF Model:	4653.87	0.36	-68.54	0.68
shakespeare-hamlet-25.txt	33098.00	4842.00	-1.20	MMF Model:	1674.85	2.35	-93.73	0.75

Table 5.4: *Documents where the rank & frequency relationship followed MMF model*

The next family was the exponential models. These models have the exponential or logarithmic functions involved. They are generally convex or concave curves, but some models in this group are able to have an inflection point and a maximum or minimum. Only two documents fall under this category. The fit parameters for these documents are given below:

Vapor Pressure Model, $y = e^{(a + \frac{b}{x} + c \log(x))}$

File Name	No- of Words	Unique words	Zipf's Coeff.	Relationship between Rank & Frequency	a	b	c
lucretius-on-395.txt	75386	7446	-1.23	Vapor Pressure Model	9.04	-0.62	-1.01
365 foriegn dishes.txt	27891	1581	-1.22	Vapor Pressure Model	8.76	-1.37	-1.17

Table 5.5: *Documents where the rank & frequency relationship followed Vapor Pressure model*

The last family of distributions was the Power Family that involves raising one or more parameters to the power of the independent variable, or raising the dependent variable to the power of a given parameter. This family is generally a set of convex or concave curves with no inflection points or maxima/minima. Nine documents out of thirty one fall under this category. The fit parameter for these distributions is given below:

Hoerl Model, $y = ab^x x^c$

File Name	No- of Words	Unique words	Zipf's Coeff.	Relationship between Rank & Frequency	a	b	c
aristotle-meteorology-80.txt	43470.00	3186.00	-1.30	Hoerl Model	3845.7	0.99	-0.80
barrie-peter-277.txt	47885.00	4788.00	-1.29	Hoerl Model	2252.21	0.98	-0.47
twain-tom-40.txt	24486.00	2488.00	-1.24	Hoerl Model	1494.84	0.98	-0.57
hiroshima nagasaki.txt	25341.00	3448.00	-1.24	Hoerl Model	2199.15	1.00	-0.95
eidgaah.txt	4951.00	1497.00	-0.82	Hoerl Model	168.59	0.98	-0.37
urdu.txt	4035.00	424.00	-0.81	Hoerl Model	151.06	0.99	-0.51

Table 5.6: *Documents where the rank & frequency relationship followed Hoerl model*

Modified Hoerl Model, $y = ab^{\frac{1}{x}} x^c$

File Name	No- of Words	Unique words	Zipf's Coeff.	Relationship between Rank & Frequency	a	b	c
mill-subjection-217.txt	45240	4885	-1.33	Modified Hoerl Model	7004.18	0.42	-1.06
librarys.txt	37498	2809	-1.00	Modified Hoerl Model	4769.28	0.58	-1.05

Table 5.7: *Documents where the rank & frequency relationship followed Modified Hoerl model*

There was one document that adhered to Power Fit of the form $y = ax^b$. In model building there is a set of rules, but not a hard and fast rule as there is no model, which is the only best model. The value of a model lies in the efficacy with which it performs the task for which it has been constructed. Unfortunately, many functions in real world situations are nonlinear in parameters. Some nonlinear functions can be linearized by transforming the independent and/or dependent variables. But we often encounter functions that cannot be linearized so the problem of estimating the nonlinear parameter arises. This paper discusses the approaches followed in nonlinear curve fitting. Zipf's Mandelbrot approach is based on Shifted power distribution. This distribution can be linearized and this is the way one finds the Zipf's coefficient.

The aim of curve fitting in this section was to highlight the similar nature of documents. We have been partially successful in doing this. We could classify the documents in groups that show similar fits. This implies that these are the documents that are similar as far as the distribution of rank and frequency are concerned. It is clearly visible with respect to the Zipf's coefficients.

Robustness of Zipf's Law

This section was devoted to the second objective that whether the Zipf's law can be applied in these literatures.

According to Wikipedia[2], "Robustness is the quality of being able to withstand stresses, pressures, or changes in procedure or circumstance. A system, organism or design may be said to be "robust" if it is capable of coping well with variations (sometimes unpredictable variations) in its operating environment with minimal damage, alteration or loss of functionality". In statistical terms, a robust statistical test is one that performs well even if its assumptions are violated by the true model from which the data were generated.

Kawamura & Hatano[3] (2002) introduced a simple and generic model that reproduces Zipf's law. They used logarithmic scale to address the time evolution of the model as a random walk and explained how the model reproduces Zipf's law. The explanation shows that the behavior of the model is very robust and universal. According to Knudsen[4] (2001), "Zipf's law for cities is one of the most conspicuous and robust empirical facts in the social sciences". According to Marsili et al.[5] (1998), "Zipf, half a century ago, found that city sizes obey an astonishingly simple distribution law, which is attributed to the more generic *least effort principle of human behavior*.... While individuals interact, they make a compromise of preference. Somehow the ensuing compromise results in a robust statistical distribution, Zipf's law". According to Levitin[6] (2003), "One may conclude that the ubiquitous appearance of Zipf's law is based on two independent effects. The first is the fact that very general transition probabilities lead to Zipf's law. The second reason why Zipf's law is found so often is probably based on the ranking procedure, which makes Zipf structures empirically observable because they are robust under its application". Ferrer-i-Cancho & Sole[7] (2001) commented that Zipf's law has been a popular achievement of quantitative linguistics. Zipf's appears to be robust. Many models of syntactic communication assume this law. It is an obvious ingredient for any theory of language evolution. A complete theory of language requires a theoretical understanding of its implicit statistical regularities.

We had taken 31 documents spanning time, typology and language. We had chosen a sample that is quite varied. It contains some documents that are probable outliers. Outliers in the sense that they are not defined as documents. They are basically a collection of words arranged in some meaningful order. For example, the English German Business dictionary, where the words are not bound in any consequential manner but they are there because of the alphabet they begin with. Similar is the case of the Sanskrit text, where the words pertain to the synonyms of the name of lord Vishnu. Let us revisit how Zipf's law performed in these documents

File Name	No- of Words	Zipf's Coefficient
franklin-autobiography-244.txt	68157	-1.37
wollstonecraft-maria-196.txt	45874	-1.37
jefferson-autobiography-73.txt	40648	-1.37
augustine-confessions-276.txt	176014	-1.35
chaucer-canterbury-102.txt	99403	-1.35
Arabian nights entertainments.txt	90768	-1.33
mill-subjection-217.txt	45240	-1.33
freud-young-763.txt	72133	-1.3
aristotle-meteorology-80.txt	43470	-1.3
locke-concerning-111.txt	53786	-1.29
barrie-peter-277.txt	47885	-1.29
berkeley-treatise-177.txt	36342	-1.29
bunyan-pilgrims-304.txt	57122	-1.28
anonymous-beowulf-543.txt	27129	-1.27
dickens-christmas-125.txt	21818	-1.25
hiroshima nagasaki.txt	25341	-1.24
twain-tom-40.txt	24486	-1.24
lucretius-on-395.txt	75386	-1.23
keats-endymion-484.txt	31962	-1.23
365 foriegn dishes.txt	27891	-1.22
The arctic queen.txt	16703	-1.21
shakespeare-hamlet-25.txt	33098	-1.2
shakespeare-romeo-48.txt	26784	-1.15
librarys.txt	37498	-1
aladdin ger.txt	17686	-0.92
aladdin eng.txt	5319	-0.92
eidgaah.txt	4951	-0.82
urdu.txt	4035	-0.81
Eng-ger-busDictionary.txt	10089	-0.66
Eng-ger-busDictionary.txt	9107	-0.63
sanskritwork.txt	1411	-0.54

Table 5.8: *Zipf's coefficients of various documents*

The mean Zipf's coefficient was found to be -1.153 with a standard deviation of 0.2388. As envisaged, Zipf's law is pretty robust across the documents, but for the ones

mentioned above. The reasons for this could be elaborated in the following discussion. For this we have to go into the genesis of the typology of these documents. Let us begin our discussion on the dictionary first.

Dictionary is defined as "a reference book containing an alphabetical list of words, with information given for each word, usually including meaning, pronunciation, and etymology". In other words, it is a book listing the words of a language with translations into another language (as is the case here- English to German dictionary). In bilingual dictionaries, each entry has translations of words in another language. For example, in a German-English dictionary, the entry 'kosten' has a corresponding English word, 'cost' and the entry 'gesetz' has a corresponding English word, 'law'. In many languages, words are grouped together according to their true or normal origin ("root"), and these roots are arranged alphabetically. So now we can say that a dictionary is a large corpus of words collected in a certain manner. But the principle of least effort is not observed here. This is the precise reason of why Zipf's law can not perform well in this type of corpus. Sanskrit text taken here is also a different type of corpus. This document is about synonyms of the names of lord Vishnu. Hence the repetition of words was not expected. The repetitions which are present are basically the connecting words or explanatory words. So there was no question of principle of least effort present here. This is reason of bad performance of Zipf's law in this document.

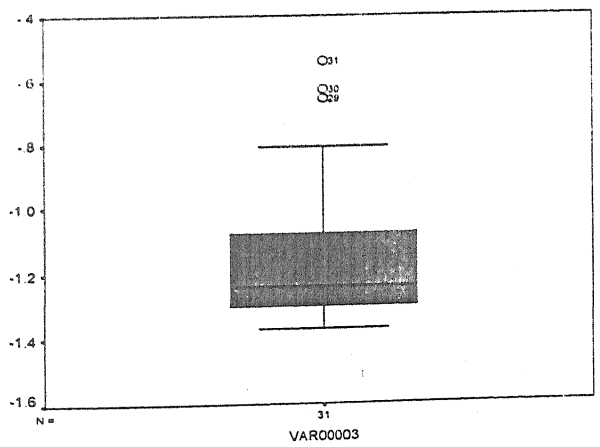


Figure 5.1: *Box plot of Zipf's coefficients of various documents*

The above figure shows a boxplot of the Zipf's coefficients. Box plots are summary plots based on the median, quartiles, and extreme values. The box represents the inter-quartile range which contains the 50% of values. The whiskers are lines that extend from the box to the highest and lowest values, excluding outliers. A line across the box indicates the median. The three documents mentioned above are outliers as far as Zipf's coefficients are concerned. But for these documents, we can conclude that Zipf's law is robust across literatures.

Inter-literature comparison of the applicability of Zipf's law

This section pertains to the third objective that is to compare the applicability of Zipf's law among different documents selected. For this we have employed cluster analysis to find cluster of documents that are "similar".

Literal meaning of clustering is to gather, to congregate or draw together. In terms of data management, clustering means dividing the data in such a way that similar data points come together. The objective of clustering is form groups that are heterogeneous but homogeneous within. Clustering is thus a method to divide a database into clusters that can be used for classification purpose. However, classification segments the data into groups that are already defined. Clustering facilitates segmentation of the data into groups that are not previously defined.

This study is thus intended to make an in-depth study on various document-parameters through cluster analysis to device a tool to formulate group(s) of documents that are 'different'. The main purpose of this study is to build a sound logic and deduce how the documents can be classified into different heterogeneous groups that are homogeneous within and try to find reasons that make the group(s) 'different'.

Clustering

According to Berry and Linoff[8] (2001), "Cluster Analysis is an important human activity. Early in Childhood, one learns to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes". With the help of clustering one can segment the data into small similar regions and thus comment on the overall distribution patterns of the data. Clustering is done on the basis of a similarity measure – attributes/variables to derive the clusters so that data points in one cluster are more similar to another (homogeneous) and data points in separate clusters are less similar or dissimilar to the data points of another cluster(s) (heterogeneous) (Anderberg[9], 1973). Clustering methods are discussed in various text books (Hartigan[10], 1975, Jain & Dubes[11], 1988). These methods have various techniques and can be performed in many ways. There are a few methods that start by considering all records to be part of one big cluster and then split them into two or more smaller clusters. On the other hand, there are methods that start with each record taken as a cluster, and iteratively combine to form

clusters. The former methods are called *Divisive methods* and the latter *Agglomerative methods* (Romesburg[12], 1984, Kaufman and Rousseeuw[13], 1990). Another method is grouping of two closest objects as a single cluster and thus number of objects is reduced to $n-1$. Then next two objects are grouped and the process continues till all n objects are covered under single cluster. Here the clustering is done step-by-step and the method is known as "*hierarchical clustering*" (Romesburg[12], 1984).

For applying clustering techniques, data is arranged in two matrices, called "data matrix" and "dissimilarity matrix". While data matrix is a representation of n objects (such as students) with m attributes (such as gender, program, region, age, social status etc.), dissimilarity matrix is a collection of distances between the pair of objects. Data matrix can be shown as follows:

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1m} \\ \vdots & & \vdots & & \vdots \\ x_{k1} & \cdots & x_{kf} & \cdots & x_{km} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{nm} \end{bmatrix}$$

Dissimilarity matrix can be shown as follows:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,1) & 0 & & \\ \vdots & \vdots & \vdots & & \vdots \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

If the distances in the matrix are near to zero then the objects are highly similar or "near" to each other.

Calculation of Distances

According to Han and Kamber[14] (2001), one could come across various types of variables while clustering the data. The entire variables that we have in this section are interval scaled variables. These variables are continuous measurement of a roughly linear scale, e.g., weather temperature and weight and height etc. One can find the distances between the objects. This is called the Euclidian distance (d) and is defined as

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + ... + \left|x_{im} - x_{jm}\right|^2}$$

where, i and j are two m dimensional data objects represented by ($x_{i1}, x_{i2}\ x_{i3}\ ...\ x_{im}$) ($x_{j1}, x_{j2}, x_{j3}, ...\ x_{jm}$).

To exhibit how clustering methods help in defining the distance between documents, a dataset 31 documents was taken and was arranged on the basis of information on parameters like Zipf Coefficients, Number of sentences, Syllables per word and Word per sentence. On the basis of this we were able to get the following cluster of documents. Number of Cases in each Cluster is as follows

	Final Cluster Centers									
	1	2	3	4	5	6	7	8	9	10
Sentences	3428	4141	7139	2451	494	10610	11160	13578	22974	5299
Syllables per word	1.62	1.57	1.55	1.63	2.12	1.60	1.49	1.54	1.53	1.93
Word per sentence	6.53	7.15	8.07	6.81	7.62	7.83	6.46	7.32	7.66	6.70
Zipf's Coefficient	-1.16	-1.23	-1.31	-1.21	-.77	-1.28	-1.30	-1.35	-1.35	-1.14

Cluster	Cases	Documents
1	4	mill-subjection-217.txt, atomic bomb hiroshima nagasaki.txt, 365 foriegn dishes.txt, The arctic queen.txt
2	4	chaucer-canterbury-102.txt, Arabian nights entertainments.txt, locke-concerning-111.txt, Eng-ger-busDictionary.txt
3	3	franklin-autobiography-244.txt, barrie-peter-277.txt, keats-endymion-484.txt
4	1	freud-young-763.txt
5	4	dickens-christmas-125.txt, lucretius-on-395.txt, shakespeare-hamlet-25.txt, urdu.txt
6	2	aristotle-meteorology-80.txt, librarys.txt
7	1	sanskritwork.txt
8	1	berkeley-treatise-177.txt
9	1	aladdin ger.txt
10	10	jefferson-autobiography-73.txt, wollstonecraft-maria-196.txt, augustine-confessions-276.txt, bunyan-pilgrims-304.txt, anonymous-beowulf-543.txt, twain-tom-40.txt, shakespeare-romeo-48.txt, aladdin eng.txt, eidgaah.txt, Eng-ger-busDictionary.txt

Table 5.9: *Results of Cluster Analysis of Documents*

It is quite evident from the analysis that clustering techniques are not quite successful in segregating the documents into heterogeneous groups that are homogeneous within. So we decided to include more parameters. These parameters were least effort %, Contain% and Flesch Index., also we deleted sentences as a variable. With these 6 parameters we again proceeded for the Cluster Analysis and obtained the following results.

Final Cluster Centers

Cluster	1	2	3	4	5	6	7	8	9	10
Contain %	16.75	55.04	36.31	44.87	34.30	61.31	18.56	2.69	53.76	62.98
Flesch Index	-9.95	69.52	60.77	67.35	28.74	59.31	-82.83	-79.97	50.09	74.23
Least Effort %	13.90	10.93	9.87	17.22	30.24	7.93	15.13	7.80	11.21	6.97
Syllables	2.54	1.54	1.65	1.57	1.99	1.64	3.41	3.33	1.76	1.49
Words Per Sentence	1.75	7.19	6.32	6.63	9.80	8.98	1.57	4.99	8.14	7.04
Zipf Coefficient	-.66	-1.19	-.87	-1.24	-.82	-1.32	-.63	-.54	-1.27	-1.28

Table 5.10: *Final Cluster Centers & Document Parameters*

The meaning of above table is explained here. Suppose we take the fourth cluster. In the fourth cluster, there will be documents whose mean contain percentage would be 44.87%. They will have mean Flesch Readability Index of 67.35. The documents would be written by exerting mean effort of 17.22%. These documents would have on an average 1.52 syllable per sentence, 6.63 words per sentence and the Zipf's coefficient of these documents would be around -1.24.

Now clustering algorithm would work like this. It will find the distances of documents from these cluster centers on these parameters and classify the documents accordingly. A document would fall in a particular cluster if it is within acceptable distance from that cluster. We have used the SPSS software for doing this analysis. It finds out the cluster membership of the document on the basis of these distances and also calculates a cumulative distance which comments on the membership of a particular document. If a document is pretty far from the cluster centre then that document would have a week membership of that cluster and vice-versa.

We obtained the clusters of the documents in the following manner

File Name	Least effort %	Contain%	Flesch Index	Zipf Coeff	Syllables per word	Word per sentence	Cluster	Distance from Centre
Eng-Ger-bus Dictionary.txt (English)	13.90	16.75	-9.95	-0.66	2.54	1.75	1	0
shakespeare-hamlet.txt	13.72	55.18	70.42	-1.2	1.53	6.71	2	2.97
shakespeare-romeo.txt	13.22	53.43	71.97	-1.15	1.51	6.95	2	3.73
lucretius-on-395.txt	9.88	54.49	65.78	-1.23	1.58	7.15	2	3.93
barrie-peter-277.txt	10.00	59.51	71.56	-1.29	1.52	6.93	2	5.00
chaucer-canterbury.txt	8.91	60.25	69.18	-1.35	1.54	7.32	2	5.59
aladdin eng.txt	9.85	47.40	68.23	-0.92	1.54	8.05	2	7.87
aladdin ger.txt	9.23	40.90	57.9	-0.92	1.7	5	3	5.60
urdu.txt	10.51	31.72	63.63	-0.81	1.6	7.63	3	5.61
keats-endymion-484.txt	17.27	44.37	68.19	-1.23	1.56	6.59	4	0.98
dickens-christmas.txt	16.94	49.15	67.75	-1.25	1.56	6.61	4	4.31
anonymous-beowulf-543.txt	13.80	45.39	71.35	-1.27	1.52	6.5	4	5.29
The arctic queen.txt	20.85	40.57	62.09	-1.21	1.63	6.81	4	7.70
eidgaah.txt	30.24	34.30	28.74	-0.82	1.99	9.8	5	0
aristotle-meteorology-80.txt	7.33	62.51	60.99	-1.3	1.62	8.64	6	2.18
Arabian nights entertainments.txt	7.12	60.82	62.41	-1.33	1.61	8.51	6	3.28
locke-concerning.txt	7.75	64.47	56.56	-1.29	1.66	9.38	6	4.21
franklin-autobiography-244.txt	9.53	57.45	57.27	-1.37	1.66	9.38	6	4.67
Eng-ger-bus Dictionary.txt (German)	15.13	18.56	-82.83	0.63	3.41	1.57	7	0
sanskritwork.txt	7.80	2.69	-79.97	-0.54	3.33	4.99	8	0
jefferson-autobiography-73.txt	13.52	53.03	48.76	-1.37	1.78	7.63	9	2.81
wollstonecraft-maria-196.txt	12.95	55.19	52.75	-1.37	1.72	8.45	9	3.51
mill-subjection-217.txt	10.80	58.04	46.75	-1.33	1.79	8.86	9	5.49
hiroshima nagasaki.txt	13.61	48.02	46.75	-1.24	1.8	7.65	9	7.07
berkeley-treatise.txt	8.87	60.15	52.17	-1.29	1.72	8.83	9	7.15
librarys.txt	7.49	48.10	53.33	-1	1.73	7.44	9	7.54
bunyan-pilgrims-304.txt	7.48	63.76	73.73	-1.28	1.48	7.89	10	1.35
freud-young-763.txt	6.22	65.56	74.4	-1.3	1.49	6.46	10	2.75
365 foriegn dishes.txt	5.67	59.47	72.03	-1.22	1.52	6.3	10	4.41
augustine-confessions.txt	5.47	66.94	69.99	-1.35	1.53	7.66	10	6.02
twain-tom-40.txt	10.03	59.19	80.99	-1.24	1.41	6.87	10	8.34

Table 5.11: *Document Parameters & Final Clusters of documents*

This clustering has given us a meaningful segregation as it has classified the documents on the basis of their characteristics. The Zipf's coefficient is a testimony of this

segregation. This classification also substantiates the results obtained in previous sections. It also highlights the comparability of documents as far as the applicability of Zipf's law is concerned. The motive of this section was precisely that.

Let us discuss the above table from the point of view of applicability of Zipf's law. In cluster 1 we have only one document and that is English German Business Dictionary (German words).This document has a Zipf's coefficient of -0.66 and is a different type of corpus, which we have discussed earlier. So it ought to be in a unique cluster. Now let us come to the second cluster, it has six documents namely shakespeare-hamlet.txt, shakespeare-romeo.txt, lucretius-on-395.txt, barrie-peter-277.txt, chaucer-canterbury.txt and aladdin eng.txt. The Zipf's coefficients for the first five documents are ranging from -1.1.5 to -1.35. The only exception in this cluster is aladdin eng.txt which has a Zipf's coefficient of -0.92, but if one see the distance of this document from the cluster centre it is the highest. This means it is tending towards the third cluster which has two documents aladdin ger.txt and urdu.txt with Zipf's coefficients of -0.92 and -0.81.

Cluster 4 has four documents: keats-endymion-484.txt, dickens-christmas-125.txt, anonymous-beowulf-543.txt and the arctic queen.txt. All these documents have a universal Zipf's coefficient of about -1.20. So there is uniformity within this cluster. Cluster 5 has one document namely eidgaah.txt with a Zipf's coefficient of -0.82. We have earlier discussed about the peculiar and different nature of this document. Cluster 6 has four documents namely aristotle-meteorology-80.txt, Arabian nights entertainments.txt, locke-concerning.txt and franklin-autobiography-244.txt. These documents have a Zipf's coefficient of around -1.30.

Cluster 7 and 8 again pertain to peculiar documents namely eng-ger-bus Dictionary.txt (German words) and sanskritwork.txt respectively. These documents have Zipf's coefficients of -0.63 and -0.54. This was expected in view of the nature of these documents. Cluster 9 has six documents namely jefferson-autobiography-73.txt, wollstonecraft-maria-196.txt, mill-subjection-217.txt, hiroshima nagasaki.txt, berkeley-treatise.txt and librarys.txt. The Zipf's coefficients of these documents are in the range -1.29 to -1.37. There is one exception that is the librarys.txt which has the perfect Zipf's coefficient of -1. Again we can justify this by saying that this document is very distant

from cluster centre of cluster 9. Cluster 10 has five documents. All of them have Zipf's coefficients in the range -1.22 to -1.35, which is again very uniform. The documents which belong to this cluster are bunyan-pilgrims-304.txt, freud-young-763.txt, 365 foreign dishes.txt, augustine-confessions.txt and twain-tom-40.txt.

What we have achieved in this section is the fact that Zipf's law is applicable in not so robust manner in those documents which have peculiar parameters. The Zipf's coefficients depend on the nature of parameters that we have defined in the preceding discussion. We can also say that these parameters are successful in comparing the documents vis-à-vis applicability of Zipf's law.

References

1. Shaw. George. Bernard. Quotation at http://thinkexist.com
2. Wikipedia, the online encyclopedia, for the definition of robustness. http://en.wikipedia.org/wiki/Robustness
3. Kawamura, Kenji. & Hatano, Naomichi. (2002). Universality of Zipf's Law, *Journal of the Physical Society of Japan,* Vol. 71 No. 5, May, 2002 pp. 1211-1213.
4. Knudsen, Thorbjorn. (2001). Zipf's law for cities and beyond: The case of Denmark. *American Journal of Economics and Sociology*, Vol. 60, N0. 1.
5. Marsili, Matteo. & Zhang, Yi-Cheng. (1998). Interacting individuals leading to Zipf's Law. *Physical Review Letters,* Volume 80, Number 12, 2741-2744.
6. Levitin, Lev. B. (2003), "Zipf Law Revisited: A Model of Emergence and Manifestation", http://necsi.org/events/iccs6/papers/898457aad990551b57939f5742e1.pdf
7. Ferrer-i-Cancho, Ramon. & Sole, Ricard. V. (2001). The small world of human language. *Proc. R. Soc. Lond.* B (2001) 268, 2261-2265.
8. Berry, Michael. J. A., and Linoff, Gordon. S. (1997). *Data mining Techniques for Marketing, Sales and Customer Support*, John Wiley & Sons.
9. Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.
10. Hartigan, J. A. (1975). *Clustering Algorithms*, New York: John Wiley and Sons, 1975.
11. Jain, A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data*, Englewood Cliffs, NJ; Prentice Hall, 1988.
12. Romesburg, H. C. (1984). *Cluster analysis for researchers*, Belmont, CA: Lifetime Learning Publication.
13. Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley &Sons, 1990
14. Han, Jiawei. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*, CA: Morgan Kaufman Publishers, 2001.

Chapter 6

Summary, Conclusions & Suggestions for Future Research

Summary & Conclusions

Zipf formulated a law in 1930 that says frequency count (number of occurrence) of words in any text is inversely proportional to the rank of that word. Frequencies count of the words is the number of occurrences of the words in that text. The words are then arranged in the decreasing order of frequency so that the most frequent word gets the highest rank. Zipf, in his first thesis, "Relative Frequency: A Determinant of Phonetic Change" wrote, "Observing the speech of many hundreds of millions of people, we have demonstrated, in part actually, in part by induction, that the conspicuousness or intensity of any element of language is inversely proportionate to its frequency. Zipf's Law approximates the relationship between rank and frequency of any text. The text should consist of at least 5000 words in order for the product of $r * f$ to be reasonably constant. Zipf attributed this law as a consequence of "Principle of Least Effort". Zipf's distribution plays a central role in the modeling of human activities, particularly of the variable studied in bibliometrics and scientometrics. It is called "one of the most puzzling phenomena in bibliometrics". Zipf (1949) in his work, "Human Behavior and the principle of least effort" viewed language as a "tool" that is shaped by its "jobs" in human society. Other works of Zipf were "Selective Studies and the Principle of Relative Frequency in Language" which as published in 1932, "Psycho-Biology of Languages" which was published in 1935 and "National Unity and Disunity: The Nation as a Bio-Social Organism" which was published in 1941. In the study "Psycho-Biology of Languages" Zipf's goal was to put language study on a par with exact sciences, by use of "statistical techniques". It was an attempt to prove that the key to the explanation of all synchronic and diachronic language-phenomena has been found in a statistically estimated tendency to maintain equilibrium between size and frequency.

There has been a debate as to if Zipf's law follows a Power-law or "stretched exponential" (Weibull) or "log-normal" or "Yule distribution". There are two Zipf "laws": the rank-frequency one and the frequency-count one. There are many directions of thoughts about Zipf's Law. Some of them are like as follows:

- Zipf's law can be derived from stochastic processes

- Φ Zipf's law is Bose – Einstein form of the classical occupancy model
- Φ Zipf's law is a Negative Binomial Model
- Φ Zipf's law is a Logarithmic Series distribution
- Φ Many of the classical occupancy model can be manipulated to yeild hyperbolic distributions.
- Φ Zipf's law is a Beta function
- Φ Zipf's law is a cumulative advantaged distribution
- Φ Zipf's law is information theoretic approach to study the statistical structure
- Φ Works based on the field of quantitative linguistics is dependent on Zipf's law
- Φ Laplace's law of succession isshown to be the 'Zipfian' frrequency analogue of the Bradford Law
- Φ The Discrete Gaussian Exponential (DGX) as defined by proposed PDF reduces to Zipf's law as $\mu \rightarrow \infty$ etc.

Zipf did show that an astonishingly wide range of phenomena...exhibited distributional behavior that could be approximated by his 'Law'. But many people commented that it applies to the distribution of only social characteristics and the relative frequency of occurrence of the descriptors of science/technology did not conform to the theoretical distribution of Zipf. They argued that too much of emphasis has been placed on this result (Zipf's law). They commented that Zipf's law is theoretically elegant, but it provides only a loose fit to actual text and in practice must be modified by introduction of additional parameters. One major finding was that all the empirical distributions can be divided into two types. These are Gaussian type (G-type) and Zipfian type (Z-type).These Z-type distributions have no moments whatever. Zipf distribution of a document employs the frequencies of the words forming that particular document so the contextual similarity can be assessed by it on the basis of numerical encoding produced by the particular distribution.

Zipf's Law has plethora of applications in the modern times. Many researchers have applied Zipf's law in city populations. It is used in modeling urban growth patterns and

the Self-Organizing Economy. A model of a large-scale city formation is developed using Zipf's law. Others developed an intermittency model for urban development and modeled interacting individuals using Zipf's law.

It is claimed the Zipf's law governs many features of the Internet. Zipf's law has implications for the search strategies used in P2P networks. Web requests from a fixed user community are distributed according to Zipf's law. Zipf's law is used in Web Access Statistics and Internet traffic like caching relay for the World Wide Web. A model is proposed based on Zipf's law for software reliability analysis.

Zipf's law has applications in finance and business also. Many empirical size distributions in economics and elsewhere exhibit power-law behavior in the upper tail. Zipf's plots and the size distribution of firms are related. Zipf distribution thus characterizes firm sizes. Zipf distribution forms a microeconomic model in which individual agents interact to form productive teams.

Zipf's law is applied in many other areas like ecological systems, genomic data, earthquakes and clinical diagnosis etc. Zipf's law is applied in ascertaining importance of genes for cancer classification using micro array data. It is found that inverse power relationship between the rank order of diagnosis and the frequency of the appearance of these diagnoses exists (Zipf's Law). There are many more examples like Zipf's law in percolation, in immune system, in liquid gas phase transition of nuclei and in psychiatric ward.

There have been many applications of the law in natural languages, like English, Chinese, Voyanich manuscript and random texts etc. Universality of Zipf's law and the differences between all languages on Earth tempted researchers to think that its explanation has something to do with language. Zipf's law can be rooted in a language structuring process of coding, which adds redundancy necessary for language understanding. Zipf's Law provides a distributional foundation for models of the language learner's exposure to segments, words and constructs, and permits evaluation of learning models.

There have been many studies to study the effect of different corpus. Some of these examples include legal texts, Brown corpus of 1 million words of American English, large corpora in two languages, English and Mandarin, Chinese corpus, Eldridge's

distribution of word usage in four American newspaper articles, Brugmann's study of four plays in Plautine Latin, noun frequency in Macaulay's essay on Bacon, Russian corpus, Anthony and Cleopatra, Richard III, novels such as 'Withering Heights' by Emily Bronte; 'Sense and sensibility' by Jane Austin, Voynich manuscript, Hindi and Urdu texts, Greek corpus, French corpus, technical writing, spoken American (verbatim) and samples of adult speech etc. It is commented that Zipf's law is a reflection of a specific property of the organization of human memory, which usually operates with more frequent language units in all cases of the spontaneous use of speech.

Present work was carried out with the objective of finding the interrelationships between the rank and the frequency of a word in selected literatures; test whether the Zipf's law can be applied in these literatures; do an inter-literature comparison of the applicability of Zipf's law and attempt mathematical modelling. Few hypotheses were assumed to be true such as the principle of least effort is a universal phenomenon; all writers would follow an economy in the use of words irrespective of the language concerned and the rank-frequency distribution of words would be similar in all languages.

For inter-literature comparison of the applicability of Zipf's law, the study had selected the few sets of texts from diverse literatures. 31 sets of texts were selected from computer science literature, Hindi, English, German, Urdu, Sanskrit, a technical subject like Library Science, and a dictionary. Other sources include public domain electronic texts (e-texts) in the areas of American and English literature as well as Western philosophy. These were "classic" texts that have stood the test of time. They also encompass a huge time period- as far back as 400BC to the present. Also taken were popular e-texts like "365 Foreign Dishes", "The Arabian Nights Entertainments", "The Arctic Queen" and "The Atomic Bombings of Hiroshima and Nagasaki".

The Software for calculating the word frequency from the texts used in this work is "TextSTAT". All unique words were ranked at random according to their frequency of occurrence in a decreasing order. Different ranks were assigned to each of them according to Zipf's approach of random-ranks. Microsoft Excel has been used extensively to "sort" the data in the first place and "advanced filter" feature of the Excel is used to filter out the unique frequencies. Once the Zipfian data has been obtained for

the various files regression analysis and curve-fitting was done on this data. A linear fit was done in order to find the applicability of Zipf-Mandelbrot law. We have used various statistical packages like SPSS. Minitab and Curve Expert to carry out these analyses on the selected texts.

Major Findings

The major findings of this work are:

1. Zipf's law is applicable on random text in English language from Computer Science l terature. Random texts do follow Zipf's law; however the exponent varies from text to text. The method of random rank performs inferiorly to the maximal rank method and the tied rank method proposed by authors.
2. The distribution of words according to their length and the hits they are able to generate on the popular search engine "Google" follows Zipf's Law. It is a Zipf type distribution with exponent not close to unity (In fact it came out to be –3.51).
3. Zipf's Law is applicable in English Literature (Aladdin and the Wonder Lamp) and for the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$) the coefficient c in this case is –0.92.
4. Zipf's Law in German Literature (Aladdin und die Wunderlampe) produced a coefficient of –0.92.
5. Zipf's Law is not applicable for English-German Business Dictionary (Mr. Honey's Small Business Dictionary (English-German), the coefficient in this case is –0.66 that is not close to –1.
6. Zipf's Law is applicable in Hindi Literature (Eidgaah by Munshi Premchand), the coefficient is –0.82.
7. Zipf's Law is applicable in a text from Library Science Literature ("The Library", by Andrew Lang). The Zipf's coefficient here is perfect '-1'.
8. Zipf's Law is applicable in Urdu Literature (Bisat-e-Hyder by Hyder Zaheer Ansari Hyder.), the coefficient c in this case is –0.81.

9. Zipf's Law is not applicable in this piece of Sanskrit Literature ("Sri Vishnu Sahasranaamam"). For the Mandelbrot Zipf's law ($g(r) = a(r+b)^c$) the coefficient c in this text is –0.54

10. In an attempt to relate Flesch Readability Index and Zipf's law, we finally found that 21 documents are belonging to type V. These documents have different readability indices, belong to a different genre and belong to different time periods but have almost similar value for the Zipf's coefficient. This indicates that readability has little to do with the Zipf's coefficients.

11. We defined least effort % (the ratio (percentage) of the unique words to the total number of words in the data. It reveals the "effort" that the writer has done in communicating his ideas. The smaller the percentage the less is the effort of the writer) and contain %(the ratio (percentage) of the sum of rank frequencies of the Zipfian data of the document to the total number of words in the data. It reveals the amount by which the Zipfian data is able to capture the document. The higher the percentage the more is the better the containment of document in the Zipfian data).

12. Factor Analysis revealed that Factor1 comprising of contain% and the Flesch readability index together with the least effort% are explaining 92% of the variance of the data.

13. Multiple regression analysis to predict the dependent variable Zipf's coefficient revealed that contain % alone explains 82.3% of Zipf's coefficient.

14. The Zipf's coefficient of the documents can be predicted well by the variable contain% only and the variable least effort is not required.

15. Curve fitting was applied with partially success to highlight the similar nature of documents. We could classify the documents in groups that show similar fits. This implies that those documents which are that are similar as far as the distribution of rank and frequency (Zipf's coefficients) is concerned are classifiable with the help of Zipf's Law.

16. There were three documents out of thirty one which can be deemed as 'outliers' as far as Zipf's coefficients are concerned. But for these documents, we can conclude that Zipf's law is robust across literatures.

17. Application of Cluster Analysis helped us in going further deep. Clustering has given us a meaningful segregation as it has classified the documents on the basis of their characteristics. The Zipf's coefficient is a testimony of this segregation.

18. Zipf's law is applicable in not so robust manner in those documents which have peculiar parameters. The Zipf's coefficients depend on the nature of parameters.

Suggestions for Future Research

The present study concentrated on "A comparative study of robustness of Zipf's Law across literatures". The study was more intended to study the "between" documents comparison. Keeping this delimitation in view, a number of suggestions can be put forward for future research in the area:-

1. A more comprehensive sample of documents "within" a subject can be taken to study the inter-literature variability of applicability of Zipf's law.
2. There is a vast scope of future research on technical subjects such as Library & Information Science, Management Science and Computer Science.
3. More research is required to be conducted to study the applicability of Zipf's law in Indian Languages like Sanskrit, Urdu and Hindi.
4. Since India is a country of many regional languages. These applications may be extended to the regional languages of India like Tamil, Telgu, Awadhi, Punjabi, Bengali and Oriya.
5. More research can be conducted to study the "principle of least effort" in context of Indian regional language. This will help testing the belief about the richness of these languages.

Bibliography

Bibliography

1. Adamic, Lada. A. & Huberman, Bernardo. A. (2000). The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce*, 1, 5-12.

2. Adamic, Lada. A. & Huberman, Bernardo. A. (2002). Zipf's law and the Internet. *Glottometrics*, 3, 143-150.

3. Altmann, Gabriel. (2002). Zipfian linguistics , *Glottometrics* 3, pp 19-26, 2002

4. Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

5. Aoyama, H., Souma, W., Nagahara, Y., Okazaki, M. P., Takayasu, H., & Takayasu, M. (2000). Pareto's law for income of individuals and debt of bankrupt companies. *Fractals*. 8(3), 293-300.

6. Apostolos, Georgakis. A. and Li, H. (2003). Document distances using the Zipf distribution and a novel metric. *DML Technical Report,* Department of Applied Physics and Electronics, Umea University, Sweden

7. Arlitt, Martin. F., & Williamson, Carey. L. (1997). Internet web server: workload characterization and performance implications. *IEEE/ACM Transactions on Networking*, 5(5), 631-645.

8. Axtell, Robert. L. (2001). Zipf distribution for US firm sizes. *SCIENCE, 293*.

9. Balasubrahmanyan, V. K. & Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, 3:3, 177–228.

10. Bar-Ilan, J. (2001). Data Collection Methods on the Web for Informetric purposes- A Review & Analysis. *Scientometrics*, 50, 7-32

11. Bence, Valerie. & Oppenheim, Charles. (2004). Does Bradford-Zipf apply to business and management journals in the 2001 Research Assessment Exercise? *Journal of Information Science, 30(5)*, 469-474.

12. Berry, Michael. J. A., and Linoff, Gordon. S. (1997). *Data mining Techniques for Marketing, Sales and Customer Support*, John Wiley & Sons.

13. Bi, Zhiqiang., Faloutsos, C. & Korn, F. (2001). The DGX distribution for mining massive, skewed data. *Conference on Knowledge Discovery and Data Mining (KDD)* 2001.

14. Black, Paul. E. (2000). Zipf's Law: Definition, at http://hissa.nist.gov/dads/HTML/zipfslaw.html. Site accessed on 1/29/01.

15. Bliss, C. I. & Fisher, R. A. (1953). Fitting the Negative Binomial Distribution to Biological Data & Note on the Efficient Fitting of the Negative Binomial. *Biometrics* (2): 176-200; 1953.

16. Booth, A. D. (1967). A law of occurrences for of low frequency. *Information & Control*, 10(4): 386-393; April, 1967.

17. Brent, M. R. (1997). Toward a Unified Model of Lexical Acquisition and Lexical Access. *Journal of Psycholinguistic Research* 26:363-375.

18. Brookes, B. C. (1984). Towards Informetrics: Haitun, Laplace, Zipf, Bradford and the Alvey programme. *Journal of Documentation*, Vol. 40, no. 2, pp120-143.

19. Burgos, J. D. & Moreno-Tovar, P. (1996). Zipf-scaling behavior in the immune system, *Biosystems*, 39(3):227-232.

20. Carlson, J. M. & Doyle, J. (2000). Highly optimized tolerance: A mechanism for power laws in designed systems, *Physical Review E*, 60(2):1412-1427.

21. Chai, Kim. (1982). Retrieval Language of Social Sciences and Natural Sciences: A Statistical Investigation, *Journal of the American Society for Information Sciences*, Jan 1982; 33, 1; ABI/Inform Global, Page 3.

22. Champernowne, D. (1953). A model of income distribution, *Economic Journal*, 63:318-351.

23. Chao, Dennis. & D'haeseleer, Patrik. (2001). The distribution of Variable-Length Phatic Interjectives on the World Wide Web, *UNM Computer Science Department Tech Report* TR-CS-2001-23.

24. Chao, Dennis., D'haeseleer, Patrik. (2001). The distribution of Variable-Length Phatic Interjectives on the World Wide Web, *UNM Computer Science Department Tech Report* TR-CS-2001-23.

25. Chen, Weishing. & Wu, Tai-His. (1997). A non-homogeneous software reliability model based on Zipf's law. *The International Journal of Quality & Reliability Management*. Vol.14. Issue. 4; pp. 409

26. Chen, Ye-Sho. & Leimkuhler, Ferdinand. F. (1987). Analysis of Zipf's Law: An index approach. *Information Processing and Management.* 23(3): 171-182.

27. Choi, J. S., Kim, Kyungsik., Yoon, S. M., Chang, K. H., & Lee, C. Christopher. (2005). Zipf's Law Distributions in Korean Financial Markets. *Journal of the Korean Physical Society*, Vol. 47, No. 1, July 2005, pp. 171-173

28. Crovella, M. E., Bestavros, A. (1997). Self-similarity in World Wide Web traffic: evidence and possible causes, *IEEE/ACM Transactions on Networking*, 5(6):835-846.

29. Crowley, C. J. (1975). The Distribution of Citation to Scientific Papers: A Model, presented at *Midwest Sociological Society Meeting*, Chicago, April 1975 (unpublished).

30. Dahl, H. (1979). *Word Frequencies of Spoken American* (Verbatim).

31. Deacon, T. W. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain,* Norton & Company, New York.

32. Egghe, L. (1999). On the law of Zipf-Mandelbrot for multi-word phrases. *Journal of the American Society for Information Science;* Mar 1999; 50, 3; pg. 233

33. Fedorowicz, Jane. (1982). The Theoretical foundation of Zipf's Law and its application to the bibliographic database environment. *Journal of the American Society for Information Science*, pp. 285-293, Sep. 1982.

34. Ferrer-i-Cancho, Ramon. & Solé, Ricard. V. (2003). Least effort and the origins of scaling in human language, *PNAS* 2003; 100; 788-791; originally published online Jan 22, 2003.

35. Ferrer-i-Cancho, Ramon. (2005). Zipf's law from a communicative phase transition. In: *European Physical Journal B*, 47; 449–457.

36. Ferrer-i-Cancho, Ramon. & Sole, Ricard .V. (2002). Zipf's Law and Random Texts. *Advances in Complex Systems*, Vol. 5, No. 1 (2002) 1-6

37. Ferrer-i-Cancho, Ramon. & Sole, Ricard. V. (2001b). The small world of human language", *Proc. R. Soc. Lond.* B (2001) 268, 2261-2265.

38. Ferrer-i-Cancho, Ramon. & Sole, Richard. V. (2001a). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics* 8:165-174.

39. For downloading TextSTAT, the website http://www.niederlandistik.fu-berlin.de/textstat/software-en.html was used.

40. Francis, W. N. & Kucera, H. (1964). *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English*, for use with Digital Computers Department of Linguistics, Brown University, Providence, Rhode Island

41. Gabaix, X. (1999). Zipf's law for cities: an explanation, *Quarterly Journal of Economics*, 114:739-767.

42. Gelbukh, Alexander. & Sidorov, Grigori. (2001). Zipf and Heaps Laws Coefficients Depend on Language, *Proc. CICLing-2001*, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335.

43. Gernsbacher, M. A. ed. (1994). *Handbook of Psycholinguistics.* Academic, San Diego.

44. Glassman, Steve. (1994). A caching relay for the world wide web. In *First International World-Wide Web Conference*, pages 69-76 (May 1994).

45. Haitun, S. D. (1982). Stationary Scientometric Distributions. Part I. The different approximations. *Scientometrics, 4(1)*, 5-25. Part II. Non-Gaussian nature of scientific activities, *Scientometrics, 4(2)*, 89-104.Part III. The role of Zipf distribution, *Scientometrics, 4(3)*, 181-94.

46. Han, Jiawei. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*, CA: Morgan Kaufman Publishers, 2001.

47. Hartigan, J. A. (1975). *Clustering Algorithms*, New York: John Wiley and Sons, 1975.

48. Herdan, G. (1964). *Quantitative Linguistics*. Washington, D.C : Butterworths, 1964.

49. Hertzel, Dorothy. H. (1987). History of the development of ideas in bibliometrics. *Encyclopedia of Library & Information Sciences, Vol. 42*, Supplement (7), pp 180-219.

50. Hill, B. M. & Woodroofe, M. (1975). Stronger Firms of Zipf's Law. *Journal of American Statistical Association.* 70 (349); 212-219; 1975.

51. Hill, B. M. (1970)[b]. Rank Frequency form of Zipf's Law. *Journal of the American Statistical Association.* 69 (348): 1017-1026; 1974.

52. Hill, Bruce. M. (1970)[a]. Zipf's law and prior distributions for the composition of a population. *Journal of the American Statistical Association*, 65:1220-1232.

53. Hill, Bruce. M. (1974). Zipf's law and prior distributions for the composition of a population, *Journal of the American Statistical Association*, 65:1220-1232.

54. Hřebíček, Luděk. (2002). Zipf's law and text, *Glottometrics 3*, pp 27-38, 2002

55. http://users.dickinson.edu/~braught/courses/cs132f02/labs/lab07.html for information about Flesh Index. Site accessed on 02/06/2006.

56. http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart for information about Project Guttenberg e-text & e-books.

57. Hyams, Daniel. For information about CurveExpert 1.3, a comprehensive curve fitting system for Windows and for Curve Expert help documentation (http://curveexpert.webhop.net)

58. Ivancheva, Ludmila. E. (2001). The Non-Gaussian nature of Bibliometrics and Scientometric distributions: A new approach to interpretation. *Journal of the American Society for Information Science and Technology. 52*, 13, pg. 1100

59. Jain, A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data*, Englewood Cliffs, NJ; Prentice Hall, 1988.

60. Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley &Sons, 1990

61. Kawamura, Kenji. & Hatano, Naomichi. (2002). Universality of Zipf's Law, *Journal of the Physical Society of Japan*, Vol. 71 No. 5, May, 2002 pp. 1211-1213.

62. Knudsen, Thorbjorn. (2001). Zipf's law for cities and beyond: The case of Denmark, *American Journal of Economics and Sociology*, Vol. 60, No. 1.

63. Krugman, P. (1996). *The Self-Organizing Economy* (Blackwell, Cambridge, MA).

64. Laherrere, Jean. & Sornette, D. (1998). Stretched exponential distributions in nature and economy: 'Fat tails' with characteristic scales. *European Physical Journal* B2: 525-539.

65. Landini, G. (1997). Zipf's laws in the Voynich manuscript. http://sun1.bham.ac.uk/G.Landini/evmt/zipf.htm

66. Landini, G. (1997). Zipf's laws in the Voynich manuscript. http://sun1.bham.ac.uk/G.Landini/evmt/zipf.htm

67. Le Quan, Ha., Sicilia-Garcia E. I., Ming, J. & Smith, F. J. (2002). Extension of Zipf's Law to Words and Phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, pages 315-320, Taipei, Taiwan.

68. Lee Breslau., Pei, Cao., Li, Fan., Graham, Phillips. & Scott, Shenkar. (1999). Web Caching and Zipf-Like Distributions: Evidence and Implications, *IEEE INFOCOM*, Vol. XX. NO- Y.

69. Levins, R. (1966). *Am. Scientist*. 54:421-31

70. Levitin, Lev. B. (2003), "Zipf Law Revisited: A Model of Emergence and Manifestation", http://necsi.org/events/iccs6/papers/898457aad990551b57939f5742e1.pdf

71. Li, W. & Yang, Y. (2002). Zipf's law in importance of genes for cancer classification using micro array data, *Journal of Theoretical Biology*, 219:539-551.

72. Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6): 1842-1845.

73. Li, W. (1998). Letter to the editor: Zipf's law in the structure and Evolution of Languages, in *Complexity*, 3(5): 9-10

74. Li, Wentian. (2002). Zipf's Law Everywhere, *Glottometrics,* 5, 2002, 14-21

75. Li W, References on Zipf's Law, http://www.nslijgenetics.org/wli/zipf/index.html

76. Losee, Robert. M.(2001). Term Dependence: A Basis for Luhn and Zipf Models. *Journal of the American Society for Information Science and Technology.* 52 (12), 1019-1025, 2001

77. Ma, Y. G. (1999). Zipf's law in the liquid gas phase transition of nuclei, *European Physics Journal*, A6:367-371.

78. Makse, Hernan. A., Havlin, Shlomo., Stanley, H. Eugene. (1995). Modeling urban growth patterns, *Nature*, 377:608-612.

79. Mandelbrot, B. B. (1953). An informational theory of the statistical structure of languages, in *Communication Theory*, ed. W. Jackson (Butterworth, 1953), pp. 486-502.

80. Mandelbrot, B. B. (1961). Final note on a class of skew distribution functions: analysis and critique of a model due to H.A. Simon, *Information and Control*, *4*, 198-216.

81. Mandelbrot, B. B. (1963). New methods in statistical economics, *Journal of Political Economy*, 71:421-440.

82. Mandelbrot, B. B. (1997). *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk,* Springer-*Verlag.*

83. Mandelbrot, B. B. (1959). A note on a class of skew distribution function. Analysis and critique of a paper by H.A. Simon, *Information and Control*, *2,* 90-99.

84. Mandelbrot, B. B. (1961). Post scriptum to 'final note', *Information and Control*, 4, 300-304.

85. Manrubia, S. C., Zanette, D. H. (1998). Intermittency model for urban development, *Physical Review E*, 58:295-302.

86. Marsili, Matteo. and Zhang, Yi-Cheng. (1998). Interacting Individuals Leading to Zipf's Law. *Physical Review Letters,* Volume 80, Number 12, 2741-2744.

87. Martindale, Colin. & Konopka, Andrzej. K. (1996). Oligonucleotide frequencies in DNA follow a Yule distribution. *Computer & Chemistry,* 20(1): 35-38.

88. Martynyuk, Stanislav. (2006). Statistical Approach to the Debate on Urdu and Hindi. *The Annual of Urdu Studies.*

89. Milenkovic, Milan. (1997). *Operating System - Concepts and Design,* Second edition. Tata McGraw Hill, New Delhi.

90. Miller, G. A. & Newman, E. B. (1958). Tests of a statistical explanation of the rank-frequency relation for words in written English. *American Journal of Psychology*, 71, 209-218

91. Miller, George. A. & Chomsky, Noam. (1963). Finitary models of language users. In: Luce, Robert. D., Bush, Robert. R. & Galanter, Eugene. (Eds.), *Handbook of Mathematical Psychology*, vol. 2. New York: Wiley, 419–491.

92. Nicholls, Paul. Travis. (1987). Brief Communication: Estimation of Zipf Parameters. *Journal of the American Society for Information Science (1986-1998);* Nov 1987; 38, 6; pg. 443

93. Pareto, V. (1897). *Cours d'Economie Politique*. Rouge and Cie, Lausanne and Paris.

94. Parunak, Anita. (1979). Graphical analysis of ranked counts (of words), *Journal of the American Statistical Association,* Volume 74, No- 365.

95. Perline, Richard. (1996). Zipf's law, the central limit theorem, and the random division of the unit interval, *Physical Review E*, 54(1):220-223.

96. Pinker, S. & Bloom, P. (1990). Natural language and natural selection, *Behav. Brain Sci.* 13, 707–784.

97. Piqueira, J. R., Monteiro, L. H., de Magalhaes, T. M., Ramos, R. T., Sassi, R. B. & Cruz, E. G. (1999). Zipf's law organizes a psychiatric ward, *Journal of Theoretical Biology*, 198:439-443.

98. Powers, David. M.W. (1998). Applications and Explanations of Zipf's Law. In Powers, D. M.W. (ed.) NeMLaP3/CoNLL98: *New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 151-160.

99. Price, Derek. deSolla. (1976). A General theory of Bibliometric & other Cumulative Advantage Processes. *Journal of the American Society for Information Science*. 27(5); 292-306; Sept-Oct 1976.

100. Prun, Claudia. (1999). G.K. Zipf's conception of language as an early prototype of synergetic linguistics, *Journal of Quantitative Linguistics*, 6(1)

101. Rapaport, Anatole. (1957). The Stochastic and the 'Teleological' Rationales of Certain Distributions and the So-called Principle of Least Effort, *Behav. Sci.*. 2, 150.

102. Reed, W. J. & Jorgensen, M. (2004). The double Pareto-lognormal distribution - A new parametric model for size distribution. *Com. Stats - Theory & Methods*, Vol. 33, No. 8., 1733-1753.

103. Reed, W. J. & McKelvey, K. S. (2002). Power law behaviour and parametric models for the size-distribution of forest fires. *Ecological Modeling*, 150:239-254.

104. Reed, W. J. (2001). The Pareto, Zipf and other power laws, *Economics Letters*, 2001, vol. 74, issue 1, pages 15-19.

105. Reed, W. J. (2002). On the rank-size distribution for human settlements, *J Regional Science*, 41:1-17.

106. Reed, W. J. and Hughes B. D. (2003). On the distribution of family names. *Physica A,* 319:579-590).

107. Reed, W. J. and Hughes, B. D. (2002). On the size distribution of live genera. *J. Theor. Biol.* 217.

108. Reed, W. J., and Hughes, B. D. (2002). From gene families and genera to incomes and internet file sizes: why power-laws are so common in nature. *Phys. Rev. E* 66 067103.

109. Reed, W. J. and B. D. Hughes (2004) A model explaining the size distribution of gene and protein families, *Mathematical Biosciences*, Vol. 189, No. 1, 97-102.

110. Ridley, D. R. & Gonzales, E. A. (1994). Zipf's law extended to small samples of adult speech, *Percept. Mot. Skills*, 79:153-154.

111. Romesburg, H. C. (1984). *Cluster analysis for researchers*, Belmont, CA: Lifetime Learning Publication.

112. Rousseau, R. & Zhang, Qiaoqiao. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics*, 24(2): 201-220.

113. Rousseau, Ronald. (2001). Evolution in time of the number of hits in keyword searches on the Internet during one year, with special attention to the use of word Euro, *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, Sydney. Vol. 2, pp 619-627.

114. Rousseau, Ronald. (2002). George Kingsley Zipf: life, ideas, his law and Informetrics. *Glottometrics 3*, pp 11-18, 2002.

115. Samuelsson, C. (1996). *Relating Turing's Formula and Zipf's Law*, WVLC'96

116. Saxena, Anurag., Jauhari, Monika. & Gupta, B. M. (2007). Zipf's Law in a Random Text from English With a New Ranking Method, *DESIDOC Bulletin of Information Technology*, Vol. 27, No. 4, July 2007, pp. 51-58

117. Sen, B. K., Khong, Wye. Keen., Lee, Soo Hoon., Lim Bee. Ling., Abdullah, Mohd. Rafae,, Ting, Chang. Nguan,, Wee, Siu, Hiang. (1998). Zipf's law and writings on LIS. *Malaysian Journal of Library & Information Science*, 3(2). 93-98.

118. Shaw. George. Bernard. Quotation at **http://thinkexist.com**

119. Shi, Lei., Gu, Zhimin., Wei, Lin. and Shi, Yun. (2006). An Applicative Study of Zipf's Law on Web Cache. *International Journal of Information Technology* Vol. 12 No.4 2006

120. Sichel, H. S. (1975). On a distribution Law for word frequencies. *Journal of the American Statistical Association*. 70 (352) part I 542-547; 1975.

121. Simon, H. A. (1955). On a class of skew distribution functions, *Biometrika*, *42*:425-440.

122. Simon, H. A. (1960). Some further notes on a class of skew distribution functions, *Information and Control*, *3*, 80-88.

123. Simon, H. A. (1961). Reply to Dr. Mandelbrot's post scriptum. *Information and Control*, 4, 305-308.

124. Simon, H. A. (1961). Reply to 'final note' by Benoit Mandelbrot, *Information and Control*, 4, 217-223.

125. Sinha, R. M. K. (2007). For information about IIT Kanpur's e-text. http://www.cse.iitk.ac.in/users/langtech/anglabharti.htm

126. Situngkir, Hokky. An Observational Framework to the Zipf's Analysis among Different Languages Studies to Indonesian Ethnic Biblical Texts.

127. Smith, F. J. & Devine, K. (1985). Storing and Retrieving Word Phrases *Information Processing & Management*, Vol. 21, No. 3, pp 215-224.

128. Sornette, D., Knopoff, L., Kagan, Y. Y., Vanneste, C. (1996). Rank-ordering statistics of extreme events: application to the distribution of large earthquakes, *Journal of Geophysical Research*, 101(B6):13883-13894

129. SPSS for Windows, (2001) Release 11.0.0, Standard version, SPSS Inc.

130. Stanley, M. H. R., Buldyrev, S. V., Havlin, S., Mantegna, R. N., Salinger, M. A., Stanley, H. E. (1995). Zipf's plots and the size distribution of firms, *Economics Letters*, 49:453-457.

131. Steele, R. & Powers, D. M. W. (1998). *Evolution and Evaluation of Document Retrieval Queries.*

132. Stewart, J.A. (1994). The Poisson-Lognornal model for bibliometric/scientometric distributions, *Information Processing & Management,* Vol 30, No.2, pp 239-251.

133. Sun, Qinglam., Shaw, D. & Davis, C. H. (1999). A model for estimating the occurrence of same frequency word and the boundary between the high and low frequency words in texts, *Journal of the American Society for Information Science*, Mar 1999: 50, 3

134. Tachimori, Y. & Tahara, T. (2002). Clinical Diagnosis following Zipf's law, *Fractals,* Vol. 10 No. 3, 341-351.

135. Tague, Jean. & Nicholls, Paul. (1987). The maximal value of a Zipf size variable: Sampling properties and relationship to other parameters. *Information Processing & Management,* Vol. 23, No.2, pp 155-170.

136. The e-text from the collection of Ghazal "Bisat-e-Hyder" by Hyder Zaheer Ansari Hyder. (http://www.bisatehyder.indiaaccess.com/)

137. The e-text from the collection of Ghazal "Bisat-e-Hyder" by Hyder Zaheer Ansari Hyder. (http://www.bisatehyder.indiaaccess.com/)

138. The IIT Kanpur's e-text of roman version of "Eidgaah" by Munshi Prem Chand (http://www.munsipremchand.iitk.ac.in/auther.html)

139. The IIT Kanpur's e-text of roman version of "Eidgaah" by Munshi Prem Chand (http://www.munsipremchand.iitk.ac.in/auther.html)

140. The Project Gutenberg e-Book of "Sri Vishnu Sahasranaamam" (http://www.gutenberg.org/)

141. The Project Gutenberg e-Book of "Sri Vishnu Sahasranaamam" (http://www.gutenberg.org/)

142. The Project Gutenberg e-text of "Aladdin and the Wonder Lamp" (http://www.gutenberg.org/)

143. The Project Gutenberg e-text of "Aladdin and the Wonder Lamp" (http://www.gutenberg.org/)

144. The Project Gutenberg e-text of "Aladdin und die Wunderlampe", by Ludwig Fulda (http://www.gutenberg.org/)

145. The Project Gutenberg e-text of "Aladdin und die Wunderlampe", by Ludwig Fulda (http://www.gutenberg.org/)

146. The Project Gutenberg E-text of "Mr. Honey's Small Business Dictionary (English-German)" by Winfred Honig.

147. The Project Gutenberg E-text of "Mr. Honey's Small Business Dictionary (English-German)" by Winfred Honig.

148. The Project Gutenberg e-text of "The Library", by Andrew Lang #20 (http://www.gutenberg.org/)

149. The Project Gutenberg e-text of "The Library", by Andrew Lang #20 (http://www.gutenberg.org/)

150. The Public domain electronic texts (e-texts) in the areas of American and English literature as well as Western philosophy (http://www.infomotions.com/etexts/)

151. Thom, James. A., & Zobel, Justin. (1992). A model for word Clustering. *Journal of the American Society for Information Science, 43(9)*, 616-627

152. Turner, C. R. (1997). Relationship between vocabulary, text length, and Zipf's law, http://www.btinternet.com/~g.r.turner/ZipfDoc.htm

153. Urzua, Carlos. M. (2000). A simple and efficient test for Zipf's law. *Economics Letters.* 66, 257–260

154. Wang, C. (1989). Zipf's distribution of Chinese corpus, *Information Sciences*, 10, 1-8

155. Watanabe, M. S. (1996). Zipf's law in percolation, *Physical Review E*, 53(4):4187-4190.

156. Wikipedia, the online encyclopedia, for the definition of robustness. http://en.wikipedia.org/wiki/Robustness

157. Witten, I., & Bell, T. (1990). Source models for natural language text. *International Journal of Man-machine Studies, 32*, 545-579

158. Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law, *Library Trends*, 30, 53-64

159. Wyllys, Ronald. E. (1975). Measuring Scientific Prose with Rank-Frequency ('Zipf') curves: A New Use for an Old Phenomenon, in *Proceedings ASIS 38th Annual Meeting, Inf. Revolution, 12*, 30.

160. Zanette, D. H. and Manrubia, S. C. (1997). Role of intermittency in urban development: a model of large-scale city formation, *Physical Review Letters*, 79:523-526.

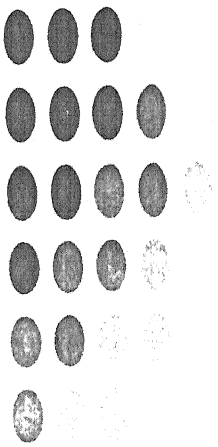
161. Zipf G.K. (1932), *Selective Studies and the Principle of Relative Frequency in Language.*

162. Zipf G.K. (1935), *Psychobiology of Languages*, Houghton-Mifflin, 1935; MIT Press.

163. Zipf, G. K. (1949). *Human Behavior and the principle of least effort,* Cambridge, MA: Addison-Wesley Press

164. Zipf, G.K. (1941). *National Unity and Disunity: The Nation as a Bio-Social Organism*, Principia Press, Bloomington Indiana, 1941.

Appendices



Appendix 1

Word Frequency distribution of 365 Foreign Dishes

w	f	w	f	w	f	w	f
and	1598	into	129	baking	67	lard	31
of	1440	sprinkle	127	remove	64	by	30
with	1050	season	125	taste	63	garlic	29
the	1018	onion	124	beat	62	celery	28
a	941	on	123	cream	61	curry	27
in	550	lemon	121	grated	60	roast	26
add	434	beaten	116	buttered	59	are	25
let	351	over	112	make	58	meat	24
until	312	fine	106	large	57	can	23
serve	311	all	103	cloves	56	be	22
to	291	pour	100	peel	55	apple	21
hot	277	fry	99	cooked	52	bits	20
cup	270	cook	98	stiff	51	do	19
salt	254	gutenberg	97	chop	50	agreement	18
then	245	or	96	fill	49	bacon	17
butter	223	cover	92	at	48	carrots	16
chopped	214	bread	91	french	46	croutons	15
pepper	202	done	90	little	44	belgian	14
some	196	tablespoonfuls	89	for	43	archive	13
eggs	185	project	87	is	42	almonds	12
boil	155	teaspoonful	82	apples	41	bavarian	11
well	151	dish	78	clean	40	access	10
mix	147	minutes	77	dough	39	about	9
brown	144	boiled	76	an	38	allspice	8
cut	142	pinch	75	glass	37	balls	7
sugar	140	thin	73	fried	36	austrian	6
bake	137	fish	71	cheese	35	anyone	5
flour	133	cold	70	out	34	additional	4
sauce	132	chicken	69	as	33	also	3
water	131	few	68	powder	32	accepted	2
						abide	1

Appendix 2

Word Frequency distribution of Aladdin and the Wonder Lamp

w	f	w	f	w	f
the	385	i	45	no	16
and	222	you	44	all	15
to	165	but	43	day	14
he	122	s	40	be	13
of	109	said	38	asked	12
aladdin	98	on	34	back	11
a	96	at	32	bade	10
his	89	who	31	am	9
him	88	not	30	an	8
her	87	so	27	about	7
in	81	as	26	africa	6
was	68	me	25	been	5
that	56	by	24	african	4
she	54	is	21	answer	3
it	52	from	20	adding	2
for	50	went	19	accident	1
had	46	son	18		

Appendix 3

Word Frequency distribution of Aladdin und die Wunderlampe

Word	Freq	Word	Freq	Word	Freq
und	561	ihr	83	am	28
der	426	vor	76	prinzessin	27
die	371	noch	75	dort	26
er	365	nach	74	gab	25
zu	262	dir	64	großvezier	24
in	219	doch	62	dieser	23
sich	209	sultan	61	ja	22
von	200	fur	59	haus	21
mit	193	nur	57	ihrem	20
sie	187	aus	55	allen	19
ich	177	durch	54	viel	18
den	164	mutter	53	diesem	17
ein	160	mich	50	ganzen	16
ihm	153	sprach	49	beim	15
nicht	147	da	47	blieb	14
dem	132	auch	46	all	13
als	126	zur	45	begann	12
das	125	bis	43	alle	11
war	117	dich	42	antwort	10
auf	112	eine	41	bau	9
du	109	denn	40	ab	8
ihn	104	nun	37	aller	7
so	99	aber	36	ach	6
zum	96	geist	35	becken	5
daß	95	schon	34	acht	4
des	92	seiner	33	abermals	3
wie	91	bei	32	a	2
an	90	dies	31	aal	1
aladdin	88	dein	30		
was	85	illustration	29		

Appendix 4

Word Frequency distribution of "The Arabian Nights Entertainments"

w	f	w	f	w	f	w	f	w	f
the	6212	s	348	man	141	another	79	daughter	38
and	3276	your	341	soon	139	make	78	back	37
to	2984	so	335	there	137	wife	77	among	36
of	2724	from	334	see	134	how	76	company	35
i	1764	one	304	been	131	long	75	brothers	34
a	1605	their	301	having	128	can	74	afterward	33
he	1302	will	298	good	127	also	73	apartment	32
in	1272	into	286	very	125	large	72	air	31
his	1062	were	270	myself	123	footnote	71	carry	30
that	982	what	267	same	122	set	70	command	29
you	979	after	253	am	119	many	69	answer	28
with	923	aladdin	236	found	118	days	68	against	27
was	849	an	222	did	113	done	67	according	26
had	787	upon	214	prince	110	asked	66	alone	25
it	769	princess	213	genie	109	heard	65	al	24
as	761	some	208	down	108	life	64	above	23
my	719	are	207	house	106	captain	62	addressed	22
me	707	sultan	205	son	104	bird	61	african	21
him	694	where	203	about	103	answered	60	approached	20
not	608	no	199	place	101	might	59	agreement	19
this	589	went	198	saw	100	merchant	58	added	18
which	586	out	196	brought	99	every	57	able	17
at	578	before	191	being	97	god	56	addressing	16
for	575	if	186	again	96	toward	54	arose	15
her	535	time	182	night	94	perceived	53	accompanied	14
but	525	do	180	gutenberg	93	opened	52	age	13
by	468	other	170	has	92	death	51	abdalla	12
they	461	day	165	should	91	morgiana	50	access	11
on	441	could	163	put	90	off	49	accepted	10
be	422	made	162	old	89	began	48	afraid	9
is	418	us	160	own	88	here	47	admire	8
she	414	came	159	ali	87	given	46	abode	7
have	409	our	156	emperor	86	both	45	aboard	6
all	407	two	155	vizier	85	away	44	accompany	5
we	387	then	154	give	84	cannot	43	abroad	4
when	378	these	152	tell	83	appeared	42	abandon	3
them	375	much	150	gave	82	dervish	41	abide	2
said	357	any	144	those	81	bahman	40	abashed	1
who	353	more	143	door	80	account	39		

Appendix 5

Word Frequency distribution of "The Arctic Queen"

w	f	w	f	w	f
the	906	me	71	an	27
and	507	it	66	e	26
of	493	s	65	did	25
to	308	at	61	deep	24
a	306	but	60	full	23
in	269	like	57	can	22
with	229	work	55	before	21
her	210	be	54	long	20
i	167	he	53	air	19
from	149	upon	52	agreement	18
this	130	eyes	50	art	17
my	128	if	48	arctic	16
that	117	love	47	away	15
or	113	we	46	back	14
on	107	thee	45	about	13
which	97	had	43	bird	12
as	95	through	41	access	11
by	94	any	39	cheeks	10
gutenberg	93	when	37	across	9
their	90	down	36	associated	8
project	87	are	35	been	7
his	82	what	33	above	6
for	80	were	32	along	5
all	76	so	31	additional	4
is	75	heart	30	accept	3
thy	74	them	29	abide	2
you	72	some	28	abandonment	1

Appendix 6

Word Frequency distribution of Meteorology by Aristotle

w	f	w	f	w	f	w	f
the	3958	bodies	156	can	74	appears	34
and	1691	some	155	either	73	after	33
of	1615	them	152	body	72	colour	32
is	1453	must	151	most	71	boiling	31
in	961	air	150	due	70	blow	30
it	885	their	144	thing	68	away	29
to	875	has	142	both	65	down	28
that	830	wind	139	another	64	contains	27
a	714	with	135	since	63	call	26
are	606	up	131	great	62	account	25
for	432	cold	130	process	61	becomes	24
as	376	like	128	cannot	60	admit	23
but	374	dry	126	where	59	actually	22
by	372	into	120	any	58	apt	21
this	369	fire	119	do	57	between	20
not	360	too	118	first	56	becoming	19
which	351	because	117	always	55	cases	18
water	349	than	115	exhalation	54	already	17
be	328	does	112	kind	53	admixture	16
we	315	those	110	called	52	appearance	15
earth	285	other	109	north	51	affected	14
when	278	being	104	about	50	affections	13
or	267	now	98	causes	49	according	12
so	261	then	96	just	48	active	11
they	257	winds	92	cloud	47	absurd	10
all	219	been	90	quantity	45	action	9
its	211	case	89	said	44	angle	8
have	194	what	88	hot	43	anaxagoras	7
sun	189	no	86	above	42	alike	6
on	188	place	85	clouds	41	able	5
same	186	was	84	day	40	absolutely	4
heat	176	will	81	change	39	accompanied	3
there	170	moisture	80	also	38	absent	2
at	167	an	79	each	37	abate	1
if	159	cause	78	explained	36		
one	157	again	77	land	35		

Appendix 7

Word Frequency distribution of "The Atomic Bombings of Hiroshima and Nagasaki"

w	f	w	f	w	f	w	f
the	2182	bomb	105	our	53	both	24
of	1225	are	102	father	51	persons	23
and	668	city	100	more	50	following	22
to	570	been	99	center	49	although	21
in	530	not	96	over	48	destruction	20
a	444	feet	94	has	47	burn	19
were	273	be	91	burns	46	come	18
was	257	have	88	fire	45	cases	17
from	251	atomic	82	cities	43	approximately	16
by	222	but	78	few	42	away	15
that	210	buildings	77	such	41	areas	14
as	176	they	76	those	39	against	13
on	172	blast	75	caused	38	again	12
at	170	up	71	bombs	37	actually	11
is	169	one	70	heat	36	aid	10
had	158	about	69	area	35	already	9
which	147	an	66	i	34	american	8
for	138	some	65	also	33	actual	7
it	133	japanese	64	almost	32	additional	6
or	126	any	63	destroyed	31	above	5
nagasaki	119	after	62	could	30	aircraft	4
this	118	out	61	distance	29	accompany	3
we	111	no	60	bombing	28	9th	2
explosion	110	pressure	56	air	27	abandoned	1
damage	108	time	55	however	26		
hiroshima	107	only	54	august	25		

Appendix 8

Word Frequency distribution- "Confessions and Enchiridion by Saint Augustine"

w	f	w	f	w	f	w	f	w	f	w	f
the	8402	will	590	may	280	spirit	170	able	91	creature	43
and	6128	one	576	earth	277	long	169	although	90	act	42
of	5416	were	570	other	276	come	168	dost	89	cast	41
to	4514	on	550	say	271	after	166	alone	88	corporeal	40
in	3516	no	536	out	270	body	165	eyes	86	abyss	39
i	3070	these	525	her	268	augustine	159	down	85	according	38
that	3040	do	520	its	263	christ	158	christian	84	authority	37
is	2810	then	503	she	262	very	154	part	83	baptism	36
it	2540	those	492	truth	257	far	149	toward	82	asked	35
not	2387	good	482	men	256	faith	148	praise	81	beautiful	34
a	2096	yet	477	didst	254	over	146	back	79	account	33
for	1979	at	470	soul	251	mercy	145	above	78	care	32
was	1667	would	468	has	249	indeed	141	apostle	77	age	31
he	1652	even	465	since	245	holy	139	times	76	catholic	30
my	1530	o	463	about	243	world	138	confess	75	anyone	29
but	1514	our	455	let	239	whole	137	beginning	74	baptized	28
as	1496	us	453	mind	236	human	135	heard	73	around	27
be	1472	their	452	know	235	make	134	born	72	added	26
this	1361	s	420	does	234	cannot	128	among	70	action	25
which	1327	because	419	said	226	upon	127	except	69	along	24
by	1292	time	418	same	222	something	126	book	68	ambrose	23
they	1120	should	379	heaven	221	against	125	having	67	acts	22
thou	1045	now	372	heart	220	death	124	books	66	afterward	21
from	1011	also	368	might	216	words	123	call	65	air	20
me	1010	could	367	before	214	called	122	different	64	aid	19
who	1008	did	366	any	209	done	120	people	63	affirm	18
we	996	how	361	therefore	206	day	118	spiritual	62	alive	17
all	992	you	357	evil	205	created	115	creator	61	abide	16
are	983	more	356	itself	203	themselves	114	actually	60	actual	15
with	978	cf	337	some	202	flesh	112	brought	59	accordingly	14
thy	952	made	329	up	200	find	111	happy	58	abroad	13
his	945	been	325	ps	197	already	110	faithful	57	abides	12
god	921	still	324	being	195	again	109	comes	56	abundance	11
what	920	an	323	art	193	certain	108	always	55	absurd	10
thee	903	life	321	many	191	hope	105	believed	54	abandoned	9
had	782	than	315	way	189	believe	103	angels	53	acquired	8
have	771	nor	313	away	188	power	102	went	52	abundant	7
when	747	chapter	311	nothing	186	sins	101	new	51	abandon	6
if	738	love	303	am	185	church	100	secret	50	ability	5
them	715	only	302	TRUE	184	order	98	became	49	absence	4
things	696	into	301	himself	181	desire	97	confessions	48	abhorred	3
so	690	through	300	light	179	creation	96	FALSE	47	abandoning	2
or	634	can	293	both	178	here	95	forever	46	ab	1
him	608	own	289	though	175	form	94	face	45		
man	595	see	286	great	174	rather	92	began	44		

Appendix 9

Word Frequency Distribution of "The Pilgrim's Progress, by John Bunyan"

w	f	w	f	w	f	w	f	w	f
the	2618	are	274	am	125	himself	69	besides	33
and	2064	chr	273	before	124	going	68	answered	32
to	1712	man	265	how	123	talk	66	behold	31
of	1554	christian	255	more	122	any	65	fall	30
that	1389	their	253	went	116	city	64	answer	29
i	1155	which	252	again	114	many	63	cast	28
he	1033	now	235	should	113	began	62	interpreter	27
in	862	thou	231	some	111	like	61	almost	26
a	808	will	230	these	107	even	59	coming	25
they	774	if	215	such	105	asked	58	ever	24
his	629	did	214	may	103	nor	57	apollyon	23
for	599	also	196	thy	99	through	56	celestial	22
but	587	when	195	hope	98	heard	55	discourse	21
him	569	on	192	our	97	day	54	above	20
is	561	do	188	let	95	death	53	according	19
it	560	from	184	very	93	being	52	along	18
you	529	go	175	back	92	about	51	bottom	17
as	511	good	174	place	90	against	50	bear	16
was	496	things	168	been	88	away	49	afraid	15
with	493	out	165	well	87	too	48	already	14
this	475	come	158	lord	85	another	47	always	13
not	431	no	155	made	84	myself	46	ask	12
be	422	came	153	could	83	hopeful	45	able	11
them	420	into	152	must	82	sin	44	agree	10
by	403	upon	148	world	81	art	43	alive	9
my	402	men	146	say	79	bid	42	abhor	8
so	398	one	145	gate	78	both	41	add	7
then	364	shall	141	can	77	cannot	40	alas	6
have	333	or	140	name	76	her	39	abide	5
me	327	god	135	an	75	company	38	abroad	4
had	301	faith	130	after	74	brother	37	abraham	3
what	287	thee	129	heart	73	best	36	abode	2
all	280	see	127	great	72	doth	35	abandoned	1
at	275	us	126	down	71	gave	34		

Appendix 10

Word Frequency distribution of Peter Pan by James M. Barrie

w	f	w	f	w	f	w	f
the	2317	would	216	will	83	told	37
and	1413	one	211	how	82	any	36
to	1197	no	203	go	79	almost	35
he	1061	are	193	more	78	called	34
a	932	what	175	came	76	also	33
was	929	hook	172	can	74	another	32
of	856	by	171	mrs	72	each	31
i	828	out	167	your	71	day	30
in	671	up	165	don	69	believe	29
that	624	t	164	come	68	best	28
she	602	if	163	let	66	great	27
they	586	now	162	nana	64	air	26
had	501	then	151	down	63	crocodile	25
you	494	who	145	back	62	asleep	24
but	481	could	141	bed	61	against	23
his	476	been	136	such	60	answer	22
her	469	do	133	last	59	girl	21
i	468	did	129	away	58	evening	20
peter	401	time	125	hand	57	already	19
not	391	which	124	long	56	cabin	18
for	379	darling	118	smee	55	better	17
wendy	358	michael	109	always	54	both	16
said	357	see	108	after	52	above	15
on	353	about	107	look	51	adventure	14
is	339	me	105	though	50	adventures	13
him	338	little	104	am	49	ago	12
as	334	into	102	really	48	across	11
at	328	boys	101	asked	47	able	10
s	285	an	98	father	46	arrow	9
them	278	children	97	every	45	ah	8
have	256	again	95	ever	44	afraid	7
all	254	like	94	home	43	added	6
so	247	know	93	get	42	admitted	5
were	244	only	90	good	41	aboard	4
be	241	night	89	has	40	abandoned	3
this	221	first	86	looking	39	abruptly	2
their	217	way	85	eyes	38	aback	1

Appendix 11

Word Frequency distribution of Beowulf from "The Harvard Classics, Volume 49"

w	f	w	f	w	f	w	f
the	1830	they	104	thou	52	band	23
of	1100	men	102	earth	51	again	22
and	719	hall	101	hand	50	doom	21
to	564	now	98	far	49	edge	20
in	490	when	97	hoard	48	atheling	19
his	443	my	95	grendel	46	any	18
he	353	no	93	have	44	are	17
that	349	but	91	could	43	against	16
with	295	beowulf	89	warrior	42	banquet	15
s	272	battle	88	came	41	along	14
a	250	er	83	god	40	about	13
was	248	king	82	foe	39	beloved	12
for	246	folk	80	hygelac	38	another	11
i	182	one	75	hardy	37	ale	10
on	178	gold	71	danes	36	bloody	9
him	173	life	70	fight	35	answer	8
from	160	me	68	home	34	aged	7
by	157	or	67	days	32	age	6
it	147	son	61	blood	31	aloft	5
as	143	man	60	after	30	abode	4
all	138	be	59	earl	29	able	3
had	135	death	58	brave	28	abide	2
this	126	them	57	an	27	abandoned	1
at	123	nor	55	barrow	26		
is	121	hero	54	hard	25		
then	109	she	53	bairn	24		

Appendix 12

Word Frequency distribution of- A Treatise Concerning "The Principles of Human Knowledge" by George Berkeley

w	f	w	f	w	f	w	f
the	1859	no	159	men	70	far	32
of	1471	with	158	therefore	69	colour	31
and	1169	things	154	say	68	had	30
to	1050	those	151	same	67	another	29
that	814	idea	149	you	66	could	28
is	773	being	148	words	65	after	27
in	716	on	147	into	63	about	26
it	693	if	143	must	62	absolute	25
or	537	only	126	like	61	common	24
be	511	other	125	cannot	60	appear	23
a	509	at	124	extension	57	concerning	22
which	425	sense	117	great	56	considered	21
are	383	their	116	its	55	clear	20
by	382	one	113	his	54	external	19
not	373	can	106	how	53	absurd	18
we	365	do	104	bodies	51	am	17
i	342	motion	102	man	50	abstracted	16
as	324	been	99	know	49	according	15
but	293	some	97	body	48	active	14
all	289	exist	93	itself	47	also	13
ideas	262	has	91	knowledge	46	able	12
for	255	abstract	88	figure	45	act	11
have	250	perceived	87	god	43	absolutely	10
they	241	existence	86	objects	42	away	9
from	227	particular	84	place	41	acknowledged	8
this	210	such	82	certain	40	absurdities	7
an	208	he	79	does	39	above	6
so	207	general	77	evident	38	abstracting	5
them	201	matter	75	because	37	accompany	4
any	198	see	74	each	36	absurdity	3
may	189	more	73	first	35	ablest	2
mind	188	though	72	conceive	34	abilities	1
there	166	said	71	answer	33		

Appendix 13

Word Frequency distribution of "The Canterbury Tales by Geoffrey Chaucer"

w	f	w	f	w	f	w	f	w	f
and	4440	may	344	out	162	am	80	cheere	37
that	2807	hem	327	how	160	world	78	felawe	36
the	2713	man	317	yet	157	deere	76	age	35
of	2651	hadde	316	myn	154	agayn	75	adoun	34
he	1945	shal	314	anon	153	lat	73	caste	33
in	1923	which	309	love	151	deeth	72	bad	32
to	1900	thy	308	kan	150	lady	71	born	31
i	1742	hath	307	seye	141	saugh	70	above	30
a	1657	now	291	after	140	speke	69	alwey	29
his	1622	been	288	nevere	138	from	68	alla	28
for	1477	thou	286	many	135	heed	67	ben	27
as	1224	god	280	myghte	134	moot	66	all	26
this	1173	at	273	noon	133	heigh	65	bee	25
was	1171	if	269	thanne	132	another	64	abyde	24
hir	1002	have	263	moore	131	ay	63	dayes	23
is	928	wolde	237	lord	127	faire	62	chambre	22
it	823	eek	233	fro	126	cam	61	cecile	21
so	726	tale	224	two	124	into	60	allone	20
but	713	what	221	any	120	arcite	59	among	19
with	697	seyde	220	gan	118	goon	58	beth	18
she	686	day	219	til	117	art	57	ale	17
al	670	han	215	tyme	113	custance	56	arrayed	16
hym	657	youre	213	evere	111	hous	55	adversitee	15
my	642	right	210	up	110	crist	54	anoon	14
by	513	thus	208	koude	108	joye	53	ageyn	13
me	471	men	203	allas	107	eyen	52	alway	12
no	464	unto	200	oon	106	doghter	51	agon	11
be	461	swich	198	er	105	manere	50	agayns	10
nat	458	alle	197	bothe	104	blood	49	agast	9
or	433	herte	188	lyf	102	body	48	afterward	8
ye	425	oure	186	seyn	99	housbonde	47	accord	7
ne	422	we	184	olde	95	knew	46	aboven	6
ful	415	an	183	folk	94	brother	45	abbot	5
ther	414	us	181	do	91	answerde	44	absence	4
they	401	every	177	see	90	best	43	abedde	3
yow	379	hise	176	also	89	creature	42	abayst	2
wel	374	o	174	forth	86	atte	41	aas	1
on	354	greet	171	wight	85	armes	40		
wol	353	thee	167	doun	84	aboute	39		
whan	349	sholde	166	knyght	82	bigan	38		

Appendix 14

Word Frequency distribution of Operating System - Concepts and Design by Milan Milenkovic

rank(ran)	g(r)	r(max)	g(rmax)	r(tied)	g(rt)	rank(ran)	g(r)	r(max)	g(rmax)	r(tied)	g(rt)
1	553	1	553	1	553	40	35	40	35	40	35
2	545	2	545	2	545	41	33	41	33	41	33
3	375	3	375	3	375	42	32	44	32	43	32
4	259	4	259	4	259	45	31	45	31	45	31
5	238	5	238	5	238	46	30	46	30	46	30
6	204	6	204	6	204	47	29	47	29	47	29
7	184	7	184	7	184	48	28	48	28	48	28
8	155	8	155	8	155	49	27	51	27	50	27
9	153	9	153	9	153	52	26	52	26	52	26
10	124	10	124	10	124	53	25	59	25	56	25
11	121	11	121	11	121	60	24	60	24	60	24
12	118	12	118	12	118	61	23	63	23	62	23
13	105	13	105	13	105	64	22	66	22	65	22
14	103	14	103	14	103	67	21	69	21	68	21
15	99	15	99	15	99	70	20	77	20	73.5	20
16	92	16	92	16	92	78	19	80	19	79	19
17	79	17	79	17	79	81	18	86	18	83.5	18
18	77	18	77	18	77	87	17	89	17	88	17
19	76	19	76	19	76	90	16	96	16	93	16
20	68	20	68	20	68	97	15	99	15	98	15
21	59	21	59	21	59	100	14	108	14	104	14
22	58	22	58	22	58	109	13	121	13	115	13
23	57	25	57	24	57	122	12	128	12	125	12
26	54	26	54	26	54	129	11	138	11	133.5	11
27	53	27	53	27	53	139	10	158	10	148.5	10
28	47	28	47	28	47	159	9	175	9	167	9
29	45	29	45	29	45	176	8	193	8	184.5	8
30	43	30	43	30	43	194	7	228	7	211	7
31	42	31	42	31	42	229	6	276	6	252.5	6
32	41	32	41	32	41	277	5	338	5	307.5	5
33	40	33	40	33	40	339	4	430	4	384.5	4
34	39	35	39	34.5	39	431	3	568	3	499.5	3
36	37	38	37	37	37	569	2	867	2	718	2
39	36	39	36	39	36	868	1	1775	1	1321.5	1

Appendix 15

Word Frequency distribution of "A Christmas Carol by Charles Dickens"

w	f	w	f	w	f	w	f
the	1194	there	100	like	51	done	22
and	822	all	97	man	48	back	21
a	531	by	96	been	47	cold	20
of	519	is	95	down	45	business	19
to	508	be	93	about	44	returned	18
in	404	so	92	little	43	after	17
it	392	their	89	are	42	bed	16
his	314	no	86	came	40	away	15
he	305	this	84	before	39	air	14
scrooge	289	ghost	82	do	38	always	13
that	258	one	81	own	37	called	12
i	236	or	78	door	36	among	11
with	206	from	77	might	35	because	10
as	173	if	75	again	34	against	9
s	170	spirit	72	being	33	afternoon	8
had	154	what	71	cried	32	answered	7
said	146	christmas	68	every	31	above	6
have	144	out	67	any	30	ancient	5
but	138	an	65	day	29	ago	4
him	136	them	63	fire	28	across	3
for	135	would	61	cratchit	27	abroad	2
on	134	very	58	another	26	abed	1
not	127	my	54	am	25		
at	122	who	53	bob	24		
its	117	time	52	come	23		

Appendix 16

Word Frequency distribution of "Mr. Honey's Small Business Dictionary" (English-German) by Winfred Honig (English words)

w	f	w	f	w	f
of	580	bank	29	financial	13
a	95	business	28	agency	12
price	60	bill	27	advertising	11
goods	57	payment	26	accounts	10
account	51	company	25	application	9
tax	50	demand	24	act	8
capital	47	job	23	accounting	7
market	42	pay	22	address	6
trade	40	agreement	21	accident	5
costs	38	free	20	abandon	4
for	35	commercial	19	ability	3
on	34	labour	18	abandonment	2
and	33	contract	17	a1	1
office	32	department	16		
letter	31	agent	15		
cash	30	acceptance	14		

Appendix 17

Word Frequency distribution of "Mr. Honey's Small Business Dictionary" (English-German) by Winfred Honig (German Words)

w	f	w	f	w	f
der	94	die	22	arbeit	9
br	43	preis	21	beschaftigung	8
in	42	ware	20	ab	7
auf	41	einen	18	auftrag	6
us	40	gesetz	16	aktien	5
des	34	angebot	15	abfindung	4
nicht	32	durch	14	abrechnung	3
eines	30	kapital	13	abandern	2
mit	25	nach	12	a	1
einer	24	etwas	11		
an	23	bank	10		

Appendix 18

Word Frequency distribution of Eidgaah By Munshi Prem Chand

w	f	w	f	w	f
hai	178	na	39	a	14
aura	103	yaha	37	ghara	13
hain	96	bhi	32	bade	12
ki	88	mohasina	31	amina	11
ke	84	kya	29	badi	10
men	80	paise	28	apani	9
se	70	hi	27	amma	8
hameid	69	gaya	24	aaj	7
ka	65	aba	23	age	6
ko	63	lekina	22	accha	5
nahin	56	kisi	21	abbajaan	4
eka	53	chimata	19	adalata	3
para	52	do	18	acakana	2
ho	48	hatha	17	abba	1
ne	47	gayi	16		
kara	43	hua	15		

Appendix 19

Word Frequency distribution of The Autobiography of Benjamin Franklin

w	f	w	f	w	f	w	f
the	3454	our	237	about	84	against	38
and	2373	s	235	those	82	am	37
of	2352	or	233	did	81	both	36
to	2239	who	214	your	80	accordingly	35
i	1543	when	204	may	78	began	34
a	1430	would	196	mr	77	always	33
in	1347	time	191	than	75	above	32
d	1100	is	188	another	74	affairs	31
was	955	more	183	said	73	different	30
that	895	been	181	each	72	act	29
it	785	there	161	do	71	afterwards	28
he	749	great	159	every	70	advantage	27
my	704	you	158	however	69	deal	26
for	696	good	154	assembly	68	acquaintance	25
with	631	should	153	among	67	able	24
as	627	no	146	friends	65	case	23
had	579	if	144	men	64	called	22
his	569	made	142	went	63	acquainted	21
be	497	might	141	its	62	attention	20
by	463	into	132	life	61	age	19
which	441	little	129	england	60	avoid	18
not	427	any	128	day	59	advice	17
but	424	could	126	father	57	arriv	16
at	423	other	124	came	56	attended	15
on	380	us	122	same	55	accounts	14
were	330	business	121	has	54	already	13
this	321	such	120	etc	53	academy	12
they	320	having	112	few	52	allow	11
we	306	before	110	must	50	acquir	10
them	296	after	109	own	49	accompanied	9
him	289	up	103	better	48	absence	8
their	282	found	102	company	47	abroad	7
from	276	out	101	account	46	abilities	6
have	272	upon	99	left	45	accepted	5
one	271	now	97	others	44	abbe	4
some	263	are	96	long	43	absurd	3
so	253	new	93	down	42	abandoned	2
being	248	only	91	again	41	abandons	1
an	240	thought	88	between	40		
all	239	many	85	himself	39		

Appendix 20

Word Frequency distribution of "A Young Girl's Diary"
Prefaced with a Letter by Sigmund Freud

w	f	w	f	w	f	w	f	w	f
the	2273	can	337	course	137	since	78	ask	35
i	2211	father	336	more	136	simply	77	bed	34
to	2187	what	332	came	135	every	74	april	33
and	2121	day	331	school	133	give	73	able	32
she	1467	are	308	any	131	again	72	best	31
a	1392	really	300	our	129	another	71	afraid	30
that	1273	then	296	could	128	into	69	although	29
it	1262	because	276	frightfully	126	asked	68	dull	28
of	992	if	263	much	120	lesson	67	christmas	27
is	990	says	249	got	119	morrow	66	angry	26
in	949	would	246	should	118	even	65	allowed	25
s	826	only	234	first	117	long	64	anneliese	24
but	817	know	231	shall	116	girl	63	care	23
not	816	there	228	after	114	great	62	began	22
her	803	did	220	does	113	ill	61	annoyed	21
for	798	or	219	still	112	room	60	box	20
was	756	do	217	get	110	gave	59	afterwards	19
we	749	m	213	his	109	back	57	big	18
t	649	go	212	other	107	better	56	baby	17
said	631	like	208	yesterday	106	lovely	55	case	16
he	628	don	202	oswald	105	aunt	54	absurd	15
so	548	such	196	who	104	thought	53	bother	14
me	532	always	195	home	103	happened	52	allow	13
you	513	been	187	by	101	awful	50	age	12
one	508	him	184	before	100	alone	49	absolutely	11
at	489	no	181	little	97	already	48	affair	10
with	482	will	180	ada	95	december	47	account	9
have	480	going	171	write	94	over	46	actress	8
all	472	this	169	thing	92	anyone	45	across	7
hella	451	anything	168	away	91	anyhow	44	above	6
mother	446	must	166	girls	90	afternoon	43	abnormal	5
had	433	never	165	how	88	understand	42	act	4
about	394	from	161	something	87	both	41	admit	3
dora	386	very	152	good	86	certainly	40	absent	2
when	376	an	151	way	83	franke	39	aback	1
be	355	frau	150	made	81	bad	38		
as	349	awfully	148	make	80	child	37		
has	341	now	138	everything	79	looks	36		

Appendix 21

Word Frequency distribution of Autobiography
By Thomas Jefferson 1743 – 1790 (With the Declaration of Independence)

w	f	w	f	w	f
the	3157	mr	111	out	38
of	2185	one	109	do	37
to	1422	these	100	assembly	36
and	1069	so	95	england	35
in	823	no	92	after	34
a	748	him	90	each	33
that	649	us	88	among	32
it	504	people	86	having	31
was	468	other	84	also	30
by	397	de	81	must	29
on	382	are	79	agreed	28
for	381	who	78	about	27
be	378	general	77	body	26
i	355	there	75	ever	25
their	329	time	74	bill	24
which	294	government	72	become	23
had	288	therefore	68	britain	22
as	267	congress	67	accordingly	21
he	253	me	66	america	20
with	252	state	65	british	19
this	247	when	64	another	18
his	240	any	63	came	17
were	234	only	62	arrived	16
they	231	if	61	afterwards	15
our	226	without	58	altho	14
from	216	day	56	approved	13
at	204	first	55	amendment	12
not	199	could	54	able	11
would	197	great	53	already	10
should	177	has	52	abuses	9
them	165	colonies	50	add	8
we	164	made	48	action	7
all	155	new	47	abolition	6
s	141	may	46	accept	5
an	132	most	45	abolish	4
is	128	same	44	abbe	3
been	127	own	43	abandonment	2
states	121	powers	41	abandon	1
or	120	france	40		
have	112	against	39		

Appendix 22

Word Frequency distribution of "Endymion: A Poetic Romance" by John Keats

w	f	w	f	w	f	w	f
the	1198	by	135	some	58	am	25
and	1148	be	129	thus	57	cold	24
of	703	thee	126	nor	56	alone	23
a	666	was	125	did	55	aye	22
to	611	at	116	could	53	after	21
d	586	no	111	light	51	above	20
i	463	one	104	came	50	around	19
in	431	or	103	every	49	born	18
his	298	when	102	far	48	dian	17
my	290	through	101	heaven	45	alas	16
with	274	love	100	air	44	airy	15
he	264	there	97	about	43	along	14
that	259	into	91	ah	42	anon	13
for	251	its	90	down	41	another	12
s	235	had	89	before	40	caught	11
from	206	have	86	hand	39	adieu	10
all	190	who	85	may	38	amid	9
it	180	will	83	ever	37	aged	8
thou	178	where	80	green	36	anxious	7
as	177	like	77	dark	35	amber	6
not	175	she	75	been	34	across	5
so	172	our	74	gentle	33	abrupt	4
o	162	now	72	can	32	abyss	3
but	161	how	69	again	31	able	2
me	157	are	68	death	30	abate	1
this	154	yet	63	golden	29		
is	153	an	62	art	28		
on	151	sweet	61	against	27		
her	144	those	59	flowers	26		

Appendix 23

Word Frequency distribution of "The Library" by Andrew Lang

w	f	w	f	w	f
the	*2746*	some	*91*	early	*33*
of	*1913*	more	*89*	about	*32*
and	*1156*	like	*87*	also	*31*
a	*906*	been	*86*	could	*30*
in	*794*	we	*83*	almost	*29*
to	*789*	when	*82*	after	*28*
is	*540*	i	*79*	famous	*27*
his	*367*	had	*78*	among	*26*
that	*314*	m	*76*	always	*25*
s	*313*	them	*75*	bound	*24*
he	*308*	will	*74*	day	*23*
books	*303*	very	*71*	best	*22*
are	*301*	most	*67*	again	*21*
for	*300*	its	*66*	being	*20*
it	*299*	no	*65*	against	*19*
with	*279*	if	*59*	ago	*18*
as	*276*	many	*57*	artists	*17*
but	*243*	collector	*55*	back	*16*
be	*237*	first	*54*	age	*15*
which	*230*	can	*53*	beauty	*14*
was	*225*	him	*52*	able	*13*
or	*224*	even	*51*	account	*12*
by	*220*	only	*48*	already	*11*
book	*206*	own	*47*	afterwards	*10*
on	*205*	man	*46*	above	*9*
not	*203*	library	*45*	alone	*8*
at	*186*	good	*44*	absence	*7*
this	*171*	now	*43*	according	*6*
have	*160*	made	*42*	absolutely	*5*
an	*136*	illustrated	*41*	absent	*4*
one	*121*	then	*40*	accessory	*3*
from	*116*	any	*39*	abominable	*2*
there	*114*	into	*38*	ab	*1*
may	*109*	amateur	*37*		
all	*106*	collection	*36*		
de	*103*	artist	*35*		
so	*92*	copy	*34*		

Appendix 24

Word Frequency distribution of "Concerning Civil Governmen"t, Second Essay- an essay concerning the true original extent and end of Civil Government, by John Locke, Chapter I

w	f	w	f	w	f	w	f
the	3284	being	203	reason	88	act	37
of	2378	nature	202	up	87	body	36
and	2012	people	198	authority	86	born	35
to	2004	will	197	life	85	earth	34
in	951	law	191	cannot	84	amongst	33
a	936	state	187	free	78	case	32
that	934	if	178	would	77	care	31
it	864	when	177	could	76	distinct	30
is	776	other	173	must	74	civil	29
be	666	such	169	commonwealth	73	about	28
as	612	can	167	nor	72	actions	27
his	592	laws	164	take	71	again	26
for	540	was	163	same	70	due	25
he	533	there	161	liberty	69	able	24
or	524	those	154	world	68	bound	23
by	518	an	151	first	66	after	22
they	457	into	148	king	65	age	21
their	451	legislative	145	against	64	conquest	20
which	429	what	144	time	63	anybody	19
power	408	on	138	out	62	better	18
but	399	at	135	community	60	apt	17
any	387	had	133	great	59	agreement	16
not	386	over	132	thus	58	certainly	15
have	369	another	125	absolute	57	acting	14
all	348	only	124	these	56	actually	13
are	330	force	120	because	55	acts	12
them	314	themselves	119	been	53	above	11
this	313	under	114	end	51	advantage	10
one	303	made	112	never	50	ask	9
no	276	every	110	before	49	ages	8
so	262	make	109	till	48	account	7
him	257	property	106	also	47	absolutely	6
from	236	hath	102	still	46	ac	5
men	226	should	101	how	45	accountable	4
man	223	shall	96	makes	44	ability	3
may	217	consent	95	both	43	ab	2
right	215	good	94	between	42	abalienatione	1
has	213	part	93	always	41		
who	212	children	92	appeal	40		
i	208	use	91	whom	39		
government	207	do	90	cases	38		

Appendix 25

Word Frequency distribution of "On the Nature of Things" by Titus Lucretius Carus

w	f	w	f	w	f	w	f	w	f
the	4443	since	249	us	132	part	74	abroad	35
and	3329	more	248	unto	130	old	72	alone	34
of	2552	through	241	like	129	once	71	certain	33
to	1613	thus	228	fire	128	without	70	beneath	32
in	1281	our	227	down	126	back	69	blood	31
that	879	if	226	forth	125	besides	68	afar	30
with	863	now	222	such	124	every	67	affairs	29
a	824	no	220	far	123	lands	66	beyond	28
from	758	have	219	see	122	fixed	65	behold	27
all	751	earth	218	er	121	whole	64	able	26
for	637	out	217	light	120	germs	63	behind	25
by	615	those	212	hath	119	had	62	clear	24
be	587	nature	203	own	118	members	60	bones	23
as	531	yet	202	itself	117	also	59	between	22
they	526	mind	198	than	113	aught	58	bear	21
their	466	many	196	her	111	borne	57	fiery	20
it	430	these	188	seeds	109	long	56	already	19
things	415	must	184	away	108	seen	55	aloft	18
are	397	same	183	round	106	never	54	across	17
so	388	again	176	frame	102	gods	53	account	16
which	376	even	175	any	97	give	52	abounding	15
when	374	how	169	within	94	limbs	51	abide	14
but	372	each	167	come	93	has	50	becomes	13
on	359	tis	163	mighty	92	elements	49	accord	12
not	351	too	160	before	91	cannot	48	ages	11
s	344	o	159	was	90	around	47	abundant	10
nor	338	life	157	death	88	another	46	abides	9
we	335	air	154	an	87	cold	45	albeit	8
then	323	because	153	eyes	85	likewise	44	acts	7
there	315	time	151	sky	84	above	43	abodes	6
what	311	do	149	could	83	birth	42	abundance	5
at	303	some	145	winds	82	against	41	ablaze	4
its	300	man	144	naught	81	being	40	aback	3
thou	277	men	142	after	79	lest	39	abandon	2
can	266	bodies	140	sense	78	feel	38	abased	1
one	265	he	134	water	76	beasts	37		
body	253	first	133	about	75	ether	36		

Appendix 26

Word Frequency distribution of "The Subjection of Women" by John Stuart Mill

w	f	w	f	w	f	w	f	w	f
the	2932	from	224	do	107	does	53	authority	23
of	2381	if	196	will	106	first	51	business	22
to	1510	one	184	was	100	opinion	50	affairs	21
and	1250	who	183	its	99	upon	49	already	20
in	994	an	180	being	98	cannot	48	able	19
is	841	has	175	general	90	feelings	47	ages	18
a	812	no	171	she	88	against	46	above	17
it	680	those	168	case	87	always	45	bad	16
that	601	her	166	society	84	because	43	action	15
be	538	there	165	law	83	marriage	42	alone	14
which	481	only	163	nature	80	said	41	actually	13
are	442	been	155	man	79	experience	40	about	12
not	434	most	149	should	78	both	39	admit	11
by	415	than	147	such	77	almost	38	act	10
as	412	so	146	great	76	cases	37	accordingly	9
for	386	would	145	some	73	modern	36	according	8
their	381	at	140	these	72	feeling	35	acts	7
women	354	he	139	had	71	far	34	acquire	6
but	328	what	137	now	68	education	33	ability	5
they	317	more	134	many	65	another	32	absolutely	4
have	312	can	123	made	64	among	31	ablest	3
or	307	power	117	character	63	different	30	abhorrence	2
all	277	s	115	influence	60	differences	29	abandon	1
them	254	i	113	between	59	better	28		
on	247	even	112	into	58	beings	27		
any	245	when	111	could	57	allowed	26		
men	235	human	109	how	55	institutions	25		
this	232	may	108	persons	54	best	24		

Appendix 27

Word Frequency distribution of Sanskrit- "Sri Vishnu Sahasranaamam"

w	f	w	f
ya	13	brahma	3
cha	8	aapnoti	2
no	7	aachaarah	1
aum	4		

Appendix 28

Word Frequency distribution of "Hamlet" by Shakespeare

w	f	w	f	w	f	w	f
the	1148	we	152	love	68	hold	30
and	970	no	143	did	65	dear	29
to	771	on	137	then	64	been	28
of	671	are	131	speak	63	doth	27
i	635	polonius	124	hath	62	both	26
you	554	all	122	must	61	away	25
a	550	by	119	an	59	comes	24
my	514	if	116	give	58	against	23
hamlet	494	or	114	man	57	done	22
in	451	good	109	make	56	before	21
it	419	thou	107	out	55	friends	20
that	407	come	106	some	54	believe	19
is	358	let	105	am	53	about	18
not	315	they	104	than	51	another	17
lord	314	now	99	much	50	better	16
his	298	gertrude	96	clown	49	answer	15
this	297	from	95	night	48	after	14
but	271	her	91	marcellus	47	air	13
with	268	how	90	mother	46	aside	12
for	252	ophelia	88	had	45	best	11
your	242	at	87	yet	44	action	10
s	238	was	86	tell	43	age	9
me	235	like	85	thus	42	back	8
he	231	most	82	play	41	adieu	7
as	229	would	80	exit	40	ah	6
be	228	well	79	look	39	alexander	5
what	218	know	77	ay	38	above	4
king	207	ll	76	god	37	aboard	3
him	197	sir	75	can	36	absolute	2
have	182	tis	73	soul	35	abate	1
d	181	enter	72	act	34		
will	169	father	71	dead	33		
do	160	first	70	again	32		
horatio	159	us	69	bernardo	31		

Appendix 29

Word Frequency distribution of "Romeo and Juliet" by Shakespeare

w	f	w	f	w	f	w	f
and	749	nurse	149	hath	63	act	27
the	684	will	148	which	62	father	26
i	659	so	145	paris	61	being	25
to	575	thee	139	one	60	both	24
a	475	he	136	am	59	away	23
of	395	his	133	how	58	lie	22
my	357	have	127	too	57	any	21
that	352	lady	117	day	56	been	20
is	350	by	116	say	55	balthasar	19
romeo	340	shall	110	art	54	ah	18
in	324	your	103	out	53	cell	17
you	295	no	102	let	51	alone	16
s	288	friar	100	montague	49	friend	15
thou	278	all	97	dead	48	above	14
me	264	ll	91	doth	47	beauty	13
not	258	do	89	such	46	bear	12
with	252	night	88	tell	45	about	11
it	227	from	86	fair	43	against	10
for	225	then	84	prince	42	ancient	9
this	217	good	83	like	41	alas	8
juliet	211	if	82	can	40	abraham	7
be	210	an	81	why	38	age	6
but	183	on	79	first	37	after	5
o	175	laurence	78	god	36	adieu	4
what	173	go	76	heart	34	abroad	3
thy	167	death	75	gone	33	able	2
d	161	man	73	ay	32	abate	1
as	157	at	70	exit	31		
capulet	154	there	69	exeunt	30		
her	152	are	68	light	29		
love	150	well	64	back	28		

Appendix 30

Word Frequency distribution of "Tom Sawyer, Detective" By Mark Twain from "The Writings of Mark Twain, Volume XX

w	f	w	f	w	f	w	f
and	749	nurse	149	hath	63	act	27
the	684	will	148	which	62	father	26
i	659	so	145	paris	61	being	25
to	575	thee	139	one	60	both	24
a	475	he	136	am	59	away	23
of	395	his	133	how	58	lie	22
my	357	have	127	too	57	any	21
that	352	lady	117	day	56	been	20
is	350	by	116	say	55	balthasar	19
romeo	340	shall	110	art	54	ah	18
in	324	your	103	out	53	cell	17
you	295	no	102	let	51	alone	16
s	288	friar	100	montague	49	friend	15
thou	278	all	97	dead	48	above	14
me	264	ll	91	doth	47	beauty	13
not	258	do	89	such	46	bear	12
with	252	night	88	tell	45	about	11
it	227	from	86	fair	43	against	10
for	225	then	84	prince	42	ancient	9
this	217	good	83	like	41	alas	8
juliet	211	if	82	can	40	abraham	7
be	210	an	81	why	38	age	6
but	183	on	79	first	37	after	5
o	175	laurence	78	god	36	adieu	4
what	173	go	76	heart	34	abroad	3
thy	167	death	75	gone	33	able	2
d	161	man	73	ay	32	abate	1
as	157	at	70	exit	31		
capulet	154	there	69	exeunt	30		
her	152	are	68	light	29		
love	150	well	64	back	28		

Appendix 31

Word Frequency distribution of "The Wrongs of Woman" by Mary Wollstonecraft

w	f	w	f	w	f	w	f
the	2412	him	173	them	74	am	32
to	1988	could	167	house	72	day	31
of	1753	when	166	made	71	became	30
and	1214	is	164	life	70	found	29
i	1094	more	157	after	69	bed	28
a	1053	this	148	seemed	67	allowed	27
my	776	they	144	then	65	another	26
her	713	would	141	now	63	because	25
in	675	been	140	being	62	away	24
was	632	who	137	myself	61	almost	23
she	479	were	136	man	59	against	22
with	465	maria	126	up	58	brother	21
that	464	what	124	ever	55	attention	20
had	459	you	114	even	54	author	19
he	435	all	109	like	53	bosom	18
me	413	their	108	must	52	better	17
not	382	only	107	every	51	always	16
his	374	no	105	while	50	air	15
for	361	so	103	before	49	added	14
as	322	are	101	darnford	48	act	13
by	315	into	99	may	46	above	12
it	304	out	96	over	45	about	11
which	298	some	95	how	44	able	10
on	294	heart	94	felt	42	advice	9
but	264	husband	86	first	41	acquired	8
from	246	any	85	eyes	40	abode	7
be	234	should	84	herself	38	abilities	6
have	228	if	81	began	37	accept	5
or	211	own	79	affection	36	abroad	4
s	192	mother	77	again	35	abhorrence	3
an	188	jemima	76	came	34	abandoned	2
at	182	child	75	friend	33	abashed	1

Appendix 32

Word Frequency distribution of "Bisat-e-Hyder" by Hyder Zaheer Ansari Hyder

w	f	w	f
hai	155	meri	22
main	98	dekh	20
ki	82	ab	19
ka	68	hoti	18
ke	60	ne	17
se	59	aur	16
yeh	49	de	15
hain	47	e	14
ko	46	dia	13
ho	44	gaya	12
gazal	43	hui	11
na	41	aye	10
nahi	38	baat	9
kya	32	aai	8
dil	31	aaj	7
jo	30	bana	6
bhi	27	ada	5
tum	26	aab	4
un	25	aa	3
tha	24	aag	2
ek	23	a	1